Second Language Acquisition Research Series

DOING REPLICATION RESEARCH IN APPLIED LINGUISTICS

Graeme Porte and Kevin McManus

ROUTLEDGE

DOING REPLICATION RESEARCH IN APPLIED LINGUISTICS

Doing Replication Research in Applied Linguistics is the only book available to specifically discuss the applied aspects of how to carry out replication studies in Applied Linguistics. This text takes the reader from seeking out a suitable study for replication, through deciding on the most valuable form of replication approach to its execution, discussion, and writing up for publication. A step-by-step decision-making approach to the activities guides the reader through the replication research process from the initial search for a target study to replicate, through the setting up, execution, analysis, and dissemination of the finished work.

Graeme Porte is Senior Lecturer in English Language and Applied Linguistics at the University of Granada, Spain. He has been Editor of the Cambridge University Press journal *Language Teaching* for 15 years and lectures and writes on quantitative research methods with a particular emphasis on research critique and replication research.

Kevin McManus is Watz Early Career Professor in Language and Linguistics, Associate Director of the Center for Language Acquisition, and Assistant Professor of Applied Linguistics at the Pennsylvania State University, USA. His research specializations include second language learning and teaching, psycholinguistics, research methodology, and replication research.

Second Language Acquisition Research Series
Susan M. Gass and Alison Mackey, Series Editors

The *Second Language Acquisition Research* series presents and explores issues bearing directly on theory construction and/or research methods in the study of second language acquisition. Its titles (both authored and edited volumes) provide thorough and timely overviews of high-interest topics, and include key discussions of existing research findings and their implications. A special emphasis of the series is reflected in the volumes dealing with specific data collection methods or instruments. Each of these volumes addresses the kinds of research questions for which the method/instrument is best suited, offers extended description of its use, and outlines the problems associated with its use. The volumes in this series will be invaluable to students and scholars alike, and perfect for use in courses on research methodology and in individual research.

Doing Replication Research in Applied Linguistics
Graeme Porte and Kevin McManus

For more information about this series, please visit:
www.routledge.com/Second-Language-Acquisition-Research-Series/book-series/LEASLARS

Of related interest:
Second Language Acquisition
An Introductory Course, Fourth Edition
Susan M. Gass with Jennifer Behney and Luke Plonsky

Second Language Research
Methodology and Design, Second Edition
Alison Mackey and Susan M. Gass

DOING REPLICATION RESEARCH IN APPLIED LINGUISTICS

Graeme Porte and Kevin McManus

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2019
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
A catalog record for this title has been requested

ISBN: 9781138657342 (hbk)
ISBN: 9781138657359 (pbk)
ISBN: 9781315621395 (ebk)

Typeset in Bembo
by Swales & Willis Ltd, Exeter, Devon, UK

Visit the eResource at www.routledge.com/9781138657359

CONTENTS

ILLUSTRATIONS

Figures

Tables

ACKNOWLEDGEMENT

The authors would like to thank Luke Plonsky for the feedback and advice obtained on the original drafts of this book.

1

INTRODUCTION

Why Replication Research Matters

Over 20 years ago the astronomer Carl Sagan observed that:

> Science is more than a body of knowledge. It is a way of thinking; a way of skeptically interrogating the universe with a fine understanding of human fallibility. If we are not able to ask skeptical questions, to interrogate those who tell us that something is true, to be skeptical of those in authority, then, we are up for grabs for the next charlatan . . . . who comes rambling along.[1]

There is much in Sagan's words which can serve as a stimulus to all researchers, novice and seasoned alike. He reminds us that science is more than the sum of its parts. That all the facts we might have learned at school and beyond form the basic building blocks of our accumulated knowledge about a subject. However, it is then one's approach to those facts that defines the indispensable "scientific" way of looking at things. This is characterized by a questioning, critical approach to what we are told or what we read: "skeptical" is clearly a cornerstone of Sagan's attitude to scientific knowledge.

His words are particularly pertinent, however, for those interested in embarking on replication research. He calls for attention to potential error in what is claimed in that knowledge, recalling our "human fallibility". An implication here is that we are imperfect beings, and it is therefore natural for us to be wrong at times. Rather than dwell on that very human characteristic, Sagan focuses on the attitude we must adopt to such a "failing". We must be able to ask questions of what we hear and of what we read.

This book is about showing you how to go about honing this questioning attitude to what we read, and then acting upon any doubts, skepticisms, or just plain inquisitiveness that reading may have aroused. While science research has

had its share of "charlatans", we begin by taking all the research we read at face value. Our principal aim in replicating research, therefore, will not be to debunk dubious claims nor sniff out potential falsehood, but rather to return to a study that interests us, "repeating it in a particular way to establish its stability in nature and eliminate the possible influence of artifacts or chance findings" (Porte, 2012, p. 4).[2]

On our way toward that aim, which makes up the core of this book, we address many conceptual and procedural elements of replication. First, however, in this introductory chapter, we respond to some fundamental questions we have heard and anticipate on this topic.

1.1 Where Does Replication Fit in to the Research Cycle?

Many of you reading this book will have participated, or be participating, in research methods courses at your institutions. The contents of Figure 1.1, therefore, will be familiar to you.

You will doubtless have learned that research is a systematic process because it follows a series of steps like those in Figure 1.1 and, in most cases, a pre-established protocol at each step. We are presented here with a cycle of work, from identifying a research area of interest, to designing a study, carrying it out, analyzing the results, and finally announcing the outcomes to an expectant world.

In turn, this stimulates the next step in the cycle, which again is often the identification of a (related) research area. And so the cycle begins again.

In such a process, our aim is to contribute with something "new" to the current state of knowledge in our area, and by so doing add more value to it.

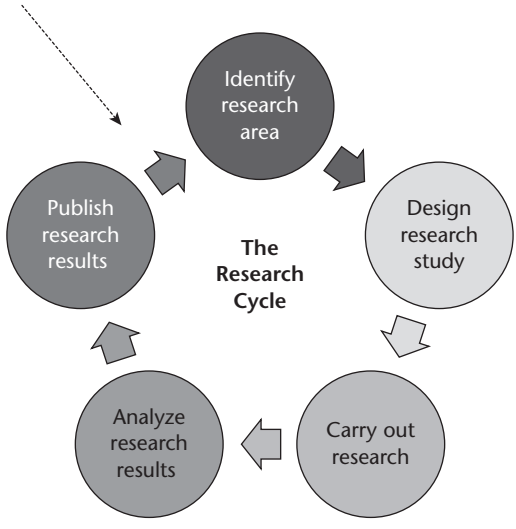


FIGURE 1.1 The research cycle.

We want to focus your attention at the point in Figure 1.1 where you see the dotted arrow. This is the critical juncture between the publication of what you have done, or what you have read, and its producing the next piece of work. This is traditionally where *you* will enter the cycle, of course. You might be encouraged by your interests, or your instructor, to produce further research at this point, using the previous literature as an appropriate springboard or stimulus for something original or innovative. Indeed, "original" and/or "innovative" is often the *sine qua non* for doctoral theses in many universities, and for many academic journals for the kind of papers they tend to publish. In other words, this next step in the cycle – and your conventional entry to it – is encouraged to be by your initiating something that follows up previous research by *extending* it into new areas and original contexts. Typically, then, you might use the previous methodology in the cycle to produce *new* data, from a *new* context with *other* participants.

Note how your *principal aim* here is not to question, nor to reconsider in any critical way the previous study's procedures and outcomes in the light of your new data. Of course, a previous study's procedures might be an important point in contextualizing your findings, and so you would want to reflect on those when your findings pattern similarly and/or differently with previous work. An important distinction between a research/extension study and a replication study, therefore, is that the extension study's principal aim is not to critically question, revisit, and/or reconsider a previous study's methodology. Thus, in an extension study, any subsequent comparison between your study and a previous one would be encouraged, but is incidental – and not your objective.

This book will have you enter the cycle at the same (dotted) point. Your intended aim, however, will shift: your contribution will come precisely from focusing on that previous step in the cycle – the publication. The stimulus for the next step and for what to research now comes from *one* earlier study, in particular, rather than a whole set of studies.

Now we need to shift our gaze back onto that previous study rather than focus just on the new one being carried out.

We now assume that you are interested in what actually went on in that previous study.

This is the stimulus for any replication study you may want to undertake. What attracted you to the study in the first place? We will be looking more closely at such stimuli in later chapters but, for now, we can suggest that you may have returned to it because it is a study everyone cites as important, or particularly significant for your area of research. Or maybe it is a study with results which somehow don't seem to "fit in" with other, similar, ones you have read. It may be that you question the generalizability (i.e., external validity) of the study's findings to other learners, contexts, or target structures. Or perhaps you just feel the results warrant closer inspection. Either way, it has piqued your interest to the point where you want to review what went on, to reconsider what emerged from it in a new light, and somehow take another look at it.

1.2 The Changing Status of Replication Research

Replicating research has sporadically appeared as a fleeting (but constant) subject of debate in the general social sciences literature since the late 1980s and early 1990s.[3] However, most of these debates remained at that level – as a discussion about whether or how it was feasible, whether it was acceptable, or even whether it was really needed. But such debate led to little uptake, and few replications were actually undertaken. Replicating previous studies as a serious research methodology has only emerged onto the applied linguistics (AL) scene relatively recently.

Many researchers busied themselves in debate about the perceived importance and feasibility of carrying out replication research in the social sciences. Meanwhile, it had long since been codified into principles for those working in the so-called "traditional" sciences as part of the skeptical approach to scientific discovery. There, the need to verify or test previous hypotheses and further probe research outcomes was long accepted as a useful means of contributing to the body of knowledge and regarded as an essential and established part of the research process. As a result, there is now an ample history of replicated studies for the researcher to consult and cite both to help better define and develop theory as well as support practice. The result is greater confidence in the interpretation and application of findings and a strong foundation upon which to base the direction of future research. It might surprise you to learn that "failures" to replicate a particular finding are equally useful, albeit any conclusions drawn from such failures will need to be accompanied by suitable caveats (see Chapter 3).

But social science deals with people, and people present us with far greater problems in the interpretation of findings from studies involving them! And the more we understand about how people behave, the more difficult social sciences become. People make up society. People are not fixed, like so many other things in the natural world. The human psyche makes the variables infinite. And so the science of studying people changes all the time, as it adapts to the different stimuli that make society and learning contexts function or change. Consequently, at best, there can only be rules of thumb, but not the types of solid physical evidence that support other (pure) sciences.

So, the argument goes – particularly in a qualitative rather than quantitative research paradigm – what we need to do in a field like AL is to accumulate as much knowledge as possible just to begin to obtain real insights into the way people go about learning languages. But accumulating data also has consequences for the way our work is consumed and interpreted. We will take this up again when we look at the reasons for replicating a study, in Chapter 2.

For now, we want you to understand that replicating a study perforce takes us on a journey back into the history of our area of interest. As we said above, we may return to a study for many reasons, but our return is predicated on the idea that no one piece of experimental research (or researcher!) can include, or control for, all the many variables that might affect an outcome. It follows that a particularly noteworthy study only stands to benefit from such renewed attention if it can have its

findings more precisely validated, its reliability focused on, its generalization tested, perhaps even delimited. Furthermore, having the benefit of a number of researchers working on the same project independently or as part of a research team can also add to the amount of detailed understanding that can be achieved. We will take up the idea of collaborative replication research work in the final chapter.

So replication matters because – whatever the outcome – a contribution to a better understanding of the target study is made. As Sagan implied, part of our task as (applied) scientists is to ask questions; in so doing, we can expect to discover error. By identifying error and having its rectification built in to our understanding of a phenomenon, we help our field progress and make our research more credible both for our fellow researchers and for practitioners in the classroom – as well as the general public.

But error won't be found if we don't look, and in AL research at least, there is evidence we are not looking. As far back as 1970 one social science researcher was already claiming that "neglecting replication is scientific irresponsibility" (Smith, 1970, p. 971).[4] Such a lack of replication over the long term can even see the academic base of the discipline brought into question. For many years, this lack of adequate evidence and relative absence of self-correction in our research was not considered of great importance. There now seems to be considerably more concern.

Fast-forward to the last decade and we witness burgeoning debate on the importance of replication research.[5]

So, are we observing here merely the periodic reappearance of the replication debate in social sciences we mentioned above? Various aspects of the current demand for replications would seem to indicate that – this time – things are more serious. The difference now appears to be that while the perceived importance of such research remains high, many of these browser hits reveal articles and blogs which carry a tone of foreboding, with many choosing to warn of "trouble at the lab", "serious problems", a "growing crisis", a "failure" in social science research, or predicting a "deepening crisis" and "a bleak outcome" for such work. What has happened in the meantime to warrant such pessimism?

» Activity

Here you see a number of papers presenting recent key discussions on the need for replication research in the social sciences. Read each carefully and take notes on:

a. What criticisms are made about the way research is currently being carried out and disseminated?
b. What does each writer see as the way forward for such research?

(continued)

(continued)

1. "Unreliable research: Trouble at the lab", October 2013, *The Economist*. www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.
2. "Errors riddled 2015 study showing replication crisis in psychology research, scientists say", March 2016, *Washington Post*. www.washingtonpost.com/news/speaking-of-science/wp/2016/03/03/errors-riddled-2015-study-showing-replication-crisis-in-psychology-research-scientists-say/.
3. "Psychology tests are failing the replication test – for good reason", August 2015, *The Guardian*. www.theguardian.com/commentisfree/2015/aug/28/psychology-experiments-failing-replication-test-findings-science.
4. "Psychology's reproducibility problem is exaggerated – say psychologists", March 2016, *Nature*. www.nature.com/news/psychology-s-reproducibility-problem-is-exaggerated-say-psychologists-1.19498.
5. "Psychology's replication crisis can't be wished away", March 2016, *The Atlantic*. www.theatlantic.com/science/archive/2016/03/psychologys-replication-crisis-cant-be-wished-away/472272/.
6. "Why does the replication crisis seem worse in Psychology?", October 2016, *Science*. www.slate.com/articles/health_and_science/science/2016/10/why_the_replication_crisis_seems_worse_in_psychology.html.

This last reading activity will have revealed one of the reasons why replication suddenly came to the attention of the general public once again. A quick search for the term "reproducibility crisis" will bring up large numbers of recent articles and discussion. The discussion appears to have centered on how much confidence we can have in the scientific research we read – and therefore to what extent apply it to learning situations – if so much appears not to have been adequately replicated or, in some cases, has proved impossible to reproduce.

This book provides instruction on how you can carry out a replication. Here is not the place to debate the relative merits of aiming for reproducibility versus replication. In theory, however, when we speak of "reproducibility" we are referring to the extent to which an outcome can be confirmed using the same approach, participants, method, and analysis. Reproducibility also encompasses research and reporting practices associated with "open science" and that can facilitate efforts to replicate and to compare the results of replication and initial studies. "Replication", by contrast, is usually understood here to involve obtaining the same outcome with variations on the original approach. However, there is debate even about the definition of "reproducible", and the confusion has led to a lack of clarity between what "reproducing" and "replicating" a study entails. As we will discuss in Chapter 4, "reproducibility" in its strictest

sense is exactly reproducing a previous study to verify its findings or add more knowledge about its generalizability, for example. We refer to this as "exact" replication and suggest such an endeavor is impossible in the field of AL (and many others!). Our use of "replication" embraces a series of modified repetitions of the original experiment along a continuum and promises equally useful contributions to the field.

The above reading activity could just as easily imply that many who depend on our research for its possible pedagogical implications and applications may be rightly concerned about the presence of undetected error or the lack of confirmatory evidence provided. Ironically, of course, one of the first lessons you learn as an experimental researcher and from any research we choose to do is that – despite all our built-in safeguards and the precision with which we prepare and execute our study, and then analyse our data – what we do will *inevitably* be subject to potentially significant limitations, bias, and error. It is all part of researching.

Now, if we admit this from the start, we also need to appreciate the consequences: our research – since it contains unavoidable error – can never be the final say on something. There will always be something else that needs to be clarified, tweaked, or further investigated as a result of our work. Maybe our choice of participants was restricted by the learning context where we found them. Maybe the randomness we wanted in selection just was not available there, or at that time. Maybe there was some confusion in the instructions given the participants, or a more powerful statistical procedure may have revealed more detailed information on the data. Maybe the presence of a relationship or the effect of an instructional treatment is clear but we are unsure of its strength.

In short, it is inevitable that some acknowledged – or as yet undiscovered – limitation will raise further questions that encourage us to go back to the study and find out something more about it.

Now, all this means we are encouraged by such a research process to move forward *and* look backward to achieve our goal. It also follows that discovery through research is not always in the one direction everybody else appears to be heading, but ironically can also be *behind* us! If we all head or experiment in the same (frontward) direction, it becomes all too easy to look for (and report) positive results only that confirm that we are all heading the right way! Replicating previous research fulfills this need to look back at and review what has led up to our present state of knowledge.

This book seeks to answer a number of questions on the practical aspects of doing replication research in AL. In particular

what a replication study is;

how to select a suitable study for replication, and

why such a study lends itself to such an approach;

what kind of replication approach is most useful given the nature of the target study;

how to carry out the study to maximize its replicative potential;

how to write up the study to highlight its comparative core; and

where to get the work published to maximize its impact on the field.

It starts from the premise that replication research is essential to the conduct of good science and the advancement of knowledge. Albeit somewhat belatedly, there is now an obvious groundswell of interest in carrying out replication research in AL and the consequential essential contribution to the body of knowledge. Searching "applied linguistics and replication" in your browser now yields a plethora of hits. Even the mere publication of this volume and that which preceded it (Porte, 2012) – the first two of their kind in the field and perhaps in any social science – would seem to signal a new era as far as replication research is concerned. Further endorsement has come from the recent publication of detailed replication reporting guidelines by the prestigious American Psychological Association[6] together with dedicated strands in leading AL journals, including *Language Teaching* and *Studies in Second Language Acquisition*, further supported by international conference roundtables and workshops.

With an eminently practical approach, this book will answer the need for more such work by showing you how to conduct meaningful replication studies and integrate them into your scholarly habits. In an ever changing and increasingly diverse research environment, it will also answer a perceived need in the field for authoritative, practical guidance on carrying out replication research in AL.

Envisage our purpose as having much more to do with handling the inherent limitations placed on any research we do, and where we do it. Much of that research is carried out in educational settings. Typically, the contextual advantages – and safeguards – afforded the researcher working in other fields are not just available. Randomization of participants, for example, is unlikely to be offered in the typical intact class setup found in schools. Treatment conditions whereby participants are presented with distinct teaching methodologies or material are unlikely to be welcomed by local authorities (or parents!). Thus often the researcher in these settings cannot determine, or take account of, the many peripheral variables which might have affected the outcome. And so we are typically presented with a study in which the investigation starts *after* the fact has occurred without interference from the researcher – providing us with tantalizingly incomplete results, often ripe for replication.

In such circumstances, and faced with the need for further insights into what has transpired in a study, we encourage you in this book to undertake replication studies as a means of improving the interpretability of research, of "filling in the gaps", if you will. When a number of such replications are carried out on a study

of interest the result is at the very least a more detailed knowledge of that study and, potentially, a more comprehensive understanding of the generalizability or otherwise of its findings.

It will be in Chapter 2 where we take up further the reasons why replicating research can often be beneficial for all those involved in researching, together with applying that knowledge to AL, and presenting it as evidence for future practice and policy. Suffice to suggest here that discovery, generalization, delimitation, in common with acceptance of the inevitability of bias and error are all concomitant, and some would say desirable concomitants, of scientific research. Without such imperfect and partial knowledge, we would not need to ask Sagan's "skeptical questions".

You should be aware at this point, however, that we do not undertake a replication study because we are assuming there has been error, or poor execution of the study, or even because we suspect something deceitful has gone on. Replicating a study is not your embarking on a criminal investigation! That said, it would be fair to say that several replication research endeavors are reacted to like this by those whose studies are replicated. We will take up this point about collaboration in the research effort in the final chapter of the book.

Chapter 2 goes on to demonstrate how to begin a search for a suitable study to replicate. This is done by encouraging you to "ask the right questions" of your reading, thereby developing a critical awareness of where a particular area of interest, and by extension a specific study, might reward a revisit. The selection routes offered move from the more general consideration of main areas of study, to sub-areas, through possible topics suggested by titles of papers through to specific research studies of interest. At each stage you are shown how to seek out target studies by making use of tools, including Google Scholar search, state-of-the-art reviews, and customized calls for replication studies in journals, and specific suggestions for follow-up research found in "limitations" descriptions often found at the end of research papers.

Chapter 3 then situates the search at the level of a specific study now identified as being of interest for possible replication. You are shown how to read a paper to raise your awareness by stimulating "a stream of consciousness". This encourages you to formulate your thoughts in questions that you might reasonably ask of the author. The chapter takes you through each section of a typical research paper in the same way, encouraging you to address aspects in detail such as participant characteristics, sample size, length and nature of the intervention, specific task variables, and analysis procedures.

Having homed in on a possible study and the reasons why you might wish to see it replicated, **Chapters 4 and 5** take you through the various replication types you might now want to use. In the first of these, you will look at internal replication, by means of the routine checking of the outcomes presented in the study from the research questions or hypotheses themselves through to the

checking of assumptions in principle statistical operations including correlation, *t*-tests, and ANOVA together with the effect sizes presented. Formal internal replication checking is then encouraged using cross-validation, jackknife, and bootstrapping.

Chapter 5 takes on the central task of planning an external replication. You are first introduced to a continuum of replication research study which foresees three types of external replication – close, approximate, and conceptual – with a cumulative number of variable changes. At each stage you are encouraged in the texts and accompanying practice activities to work step-by-step from identifying a possible variable change toward justifying the following replication and hypothesizing a possible outcome.

Chapters 6 and 7 take you through the execution and writing up of the different sections of your replication study. We focus on five components of replication execution and write-up: research questions and methodology (Chapter 6), and analysis, results, discussion, and conclusions (Chapter 7). For each component, we begin with a critique of the original/target study being replicated (Bitchener and Knoch, 2010), followed by considerations for writing up your replication using two published replication studies as models.

Our **final chapter** concentrates on the all-important dissemination of your work. Recent debate about the importance of replication research in social science research has meant that a number of publication routes have now opened up where before the replication researcher might have been met with a less enthusiastic response. We take you along these routes, providing advice and practice in selecting a suitable journal for your work, getting a journal interested in your study, and collaborating with others in replication research teams. The preparation, writing up, and submission of conference papers and poster sessions are also dealt with here. Much of the recent debate on replication research which we have described in this Introduction has highlighted the importance of following basic rules of academic etiquette and, specifically, of involving the original author/s at some stage in the process of producing your study. We take this up here and consider the recent phenomenon of "replication bullying", and how to avoid it!

As this book is intended as a text/practice book, throughout the book you are presented with boxed "activities", which are designed to give you the chance to practice further what has been presented in the main body of the chapter. We will be referring you to a number of studies within the text itself and in these activities. You will therefore need online access to the following **key papers**:

1. Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, 19, 207–217.
2. Carter, M., Ferzli, M., & Wiebe E. (2004). Teaching genre to English first language adults: A study of the laboratory report. *Research in the Teaching of English*, 38.4, 395–419.

3. Eckerth, J. (2009). Negotiated interaction in the L2 classroom. *Language Teaching*, 42.1, 109–130.
4. McManus, K., & Marsden, E. (2018). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*, 40.2, 459–475.
5. Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics*, 33.1, 66–82.

Notes

1 Interview with Charlie Rose on May 27, 1996.
2 Porte, G.K. (Ed.) (2012). *Replication Research in Applied Linguistics*. Cambridge: Cambridge University Press.
3 Neuliep, J.W. (1991). *Replication Research in the Social Sciences*. Thousand Oaks: Sage Publications.
4 Smith, Jr., N.C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, 25, 970–975.
5 Freese, J., & Petersen, D. (2017). Replication in social science. *Annual Review of Sociology*, 43, 147–165; Ishiyama, J. (2015). Replication, research transparency, and journal publications: Individualism, community models, and the future of replication studies. *PS: Political Science & Politics*, 47.1, 78–83.
6 Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, 73, 1, 3–25 (see p. 17).

2

FINDING A STUDY TO REPLICATE

Background Research

2.1 Why Might We Need to Replicate a Study?

As we have seen in the Introduction, there has been little or no focus on undertaking and publishing replication studies in the social sciences in general, and in AL in particular. Consequently, there is little tradition in theory or practice which we can use to discuss the relative importance of such research. We must look to other fields, where replication is a more deep-seated and accepted element of the research process – and to science in general – for clues to why replication research is a lynchpin to the way we go about our work.

The basic assumption when replications are undertaken against a more traditional background of doing so is that – as an accepted part of the scientific process of discovering knowledge – the original researcher has been open and transparent as regards the data and methods used. It also follows that – by participating in such a tradition and knowing others may well choose to replicate all or part of your study – the original researcher will be careful in what is presented for further scrutiny and possible replication. In other words, that researcher would assume replication in some form will be a consequence of his or her endeavors. For the consumers of this research, the motto would be, in the words of Ronald Reagan's adopted maxim, "Trust, but verify".

At the heart of replication is the need to return to a previous point, "replicare" in Latin – returning, turning around, and reviewing where you have been. Allied to this, we now have this skeptical standpoint from the scientist, wherein there is to be innate doubt built in to what is observed (or read!) and a desire to seek out more evidence for what is presented as an outcome to any research. Science itself is based around this aspect of scientific method, and since we *social* scientists are also forming hypotheses, testing them, and basing conclusions off of that data, we also qualify as scientists, in the sense also that we are assumed to execute our experimental research "scientifically".

This approach to experimental research will immediately strike you as a slow, deliberate drive for knowledge. Indeed, it would often demand an apparently retrograde movement to readdress a previous study. Scientific knowledge has to be built up very slowly as people come up with hypotheses and theories, repeatedly testing them against observations of the natural world and from other researchers in their community. Crucially, however, this accumulation of research data is used to continually refine our theories based on that community's own ideas and observations. It is not an accumulation of data from a set of loosely connected "one-off" research studies. It is patient, methodical, step-by-step work on one area of common interest, a research question if you will or hypothesis agreed on by our closed community and which we will all work on in our own contexts at the same time. Such contributions then serve to construct knowledge rather than merely amass it, since further confirmation or disconfirmation may reveal more support for one hypothesis over another; it may help us revise an established outcome on the basis of the new data; it may expose the weakness or even the erroneousness of a previous assumption; it may, together with other data arising from other replications, help us form a completely new hypothesis or assumption.

But working scientifically does that to you: the staple diet of the scientific way of doing things is repetition, recursiveness, and above all a slow and careful route to discovery. The payback, however, is that the outcomes will most likely be *comparable* data amongst that community. And data which we can go on and compare across studies is data which can indeed then be satisfactorily theorized and adequate implications drawn from it.

Working recursively means we are sanctioned by the scientific method to move forward *and* backward to achieve our goal. Thus, as we now seek out a study which might need or merit replication we need to accept that this involves us in being willing and able to question what has gone on before and how the conclusions were arrived at.

As we return to a study to reexamine it with a view to replicating it, our initial search needs to be guided by some more overarching reasons for undertaking any replication study. Later we will look at more specific reasons once we focus in on a target study.

» Activity 1: READ AND DISCUSS

Here are a number of justifications for carrying out replications of previous studies. Think about each one and decide how a replication study might achieve this.

A replication study

- can help us confirm findings or disconfirm them;
- can provide further data on the reliability and/or validity of the original study;

(continued)

(continued)

- can provide further data on the generalizability of a study;
- can provide a useful check on dubious research practices, including data manipulation;
- can help identify other possible procedures and/or data analysis methods which work better in a specific context;
- is a very useful learning tool for novice researchers;
- should be carried out by (the) researcher(s) who did not participate in the original study;
- should first be carried out by the original researcher(s) prior to publication of the original study.

2.2 Focusing on a Suitable Target Study

In the previous section, we mentioned the inevitability of limitations and error in the studies we execute and read about. To a similar extent we might expect these and other studies to be beset by flaws and biases. It follows that we need many studies to be replicated to iron out such inconsistencies and limitations and provide us with a clearer picture of what is actually going on. This is part of the normal procedure in the pure sciences: we are taught not to expect the final word on anything, and replication is part of the standard error-finding procedure through which we need to pass and upon which a sound basis of theory might eventually be safely established.

But errors, flaws, or limitations won't be found unless we go out and look for them. And given the marked tendency mentioned in the Introduction for experimental research in our field to accumulate new data rather than indulge in the sound practice of analyzing previous data, it is a reasonable assumption that there is much out there that would need or merit a further look. The recent accusations and debates in the media about the lack of replication in the social sciences seem to confirm that few people are actually looking. Specifically, in our section of that world – AL and second language acquisition – where so much time, effort, and money is spent on establishing best practices, one would have thought it was in everyone's interest to revisit experimental studies – particularly those that have been key to the development of language learning theory and practice. For a number of reasons mentioned above and in the Introduction, this has not been the case, and most of the experimental research we have seen and read in our journals reports on "new" data, and in extension or follow-up studies.

Focus on the Area of Study: Asking the Right Questions

You will have gathered that revisiting, rereading, or reanalyzing a previous paper will need to involve us much more in questioning what has gone on. It will help us develop the healthy skepticism which is needed in any address of the literature and cultivate our critical awareness for a well-reasoned and detailed scientific paper.

Our starting point in our search for a useful study which will merit the kind of attention replication brings might be the current principles or assumptions about AL we have come across in our reading. While we may well have seen these repeated in much of the literature we read and while they may appear as givens, our experience of the area might lead us to question the assumption and ask how far things are exactly as described.

» Activity 2: READ AND DISCUSS

Read some current assumptions about L2 learning and think about how far your experience leads you to agree or disagree with them. What historical research have you read about which might have led to the formulation of these principles?

- *Teachers and students should use the target language rather than the L1 in the classroom.*
- *The more meaningful exposure, the more you learn.*
- *The sooner you can acquire the grammatical system of a language, the sooner you can use the language creatively.*
- *Learners who use learning strategies effectively are more successful.*
- *Students learn best by having their errors corrected immediately by their teachers.*
- *Motivation affects the time spent learning a language.*
- *Too much correction or criticism can inhibit your learning.*
- *The more the language you are learning is like one you already know, the more quickly you will learn it.*

Your experience, intuition, and/or your reading of the literature may well have led you to question the veracity or inclusiveness of some of these statements. The next step in our search could now be to focus on an area of interest suggested by these (or other) assumptions of interest, and then search out more specific papers that might merit replication.

As we now search out a suitable study for replication, remember what was discussed in the Introduction. While in the short term replicating a study will prove invaluable practice for novice researchers who are finding their way around research methodology, doing so must also be seen as an important step in advancing your own research agenda (and career!). Therefore, select your target studies for replication with an eye to helping move forward both your own research practice and interests as well as giving service to current knowledge in the field. Opting for a randomly chosen study merely to check whether the original author made a mistake is probably not a wise move.

What follows are four possible "routes" to finding your target study: rereading any experimental research which you came across in your course reading, using

an academic search engine such as Google Scholar, reading state-of-the-art reviews for critical analysis of research, and drawing on customized calls for replication – as well as published replications of specific studies.

Course Reading

You will doubtless already have read a large number of research papers, some of which will have sparked your interest in some way – perhaps there was an unusual outcome, maybe you wondered what the outcome would have been like with more participants or from different backgrounds, or possibly you thought the instructions given to the control group could have been clearer, and so on. You might even have wondered whether – given the fact that it was carried out some time ago – the same outcomes might prevail today. As we shall see below, these are all legitimate initial stimuli for you to carry out a replication study. Similarly, the authors themselves may well have suggested how their research might be carried on or revisited in the "Limitations" section often found toward the end of the published paper.

At this point, however – once you have settled on those studies that aroused this initial curiosity – you should begin to analyze their potential for replication in more practical terms, while bearing in mind the advice in the section below, Final Considerations for Replication (see also Chapter 6 on the feasibility of replication).

» Activity 3: READ AND DISCUSS

Here are three summaries from published studies in which participant involvement is described. Look at each one and make an initial decision about the feasibility of replicating such a paper *in your research context*. Reflect upon aspects such as time needed, potential costs, and participant availability.

Study 1: The researcher advertised in the university newspaper for 100 undergraduates studying a Humanities major and offered course credits as recompense.

Study 2: The researcher took on undergraduate students from the current L2 classes at the university; however, each participant was scheduled to do the questionnaire and subsequent interview individually in the researcher's office because data had to be video recorded.

Study 3: The team of three researchers recruited participants from the campus community. The treatment group was administered the novel vocabulary-learning method weekly over two years while the control group received their normal class sessions over the same period.

Below we will look at the feasibility of carrying out the replication in more detail when we critique an individual study. However, for now, and even at this early stage of selection and reading through the abstract of a potential target study, we can respond to the labor, cost, and time requirements that might be involved in carrying out the replication. Here, for example, a quick reflection on the selection of participants and a summary of procedures might lead us to question how far in Study 3 our replicating the selection and procedures would be more onerous than the other studies. The replication may be justified or a useful contribution to knowledge, but the resources required to get a research and teaching team together for a long period as well as the recruitment from such a large sample base may prove problematic for us. We might also feel that the novel and longitudinal nature of the investigation (i.e., new teaching/learning method over two years) is not something we could commit to comfortably.

Likewise, the decision between a replication of Study 2 and 1 might hinge on the fact that participant selection might need to be based on what was available in the current "L2 classes" of interest. Moreover, video recording of the interviews could be costly and would need to be voluntary, which might mean some drop-outs from the original group and, again, there is extra time to be factored in with the individual interviewing itself.

Academic Search Engines

To fix on a topic of interest you need to narrow the overarching subject area that interests you down to a small number of prospective sub-areas (Figure 2.2). One way of finding the more "popular" or "hot" sub-areas is to go to Google Scholar and type in the main branches of study: in Figure 2.1 you can see one outcome for "Applied Linguistics". An area identified as "Bilingualism" immediately emerges from the entries, and sub-areas upon which we might now focus could include "Bilingualism and mental development", "Infant bilingualism", "Degree of bilingualism and cognitive ability", "Multilingualism and multilingual education", "Bilingualism and minority language children", and "Bilingualism and biliteracy".

Google Scholar will list the titles of any book or article on the subject, which means you may be presented with a list of theoretical papers and position papers, as well as the kind of experimental research setup which is our main focus of interest here. You will need to click through the less obvious examples to find such work and can then note down the possible topic routes suggested by the titles/keywords. Google Scholar also allows you to conduct an "Advanced" search wherein you can filter out or in certain variables aspects such as recency ("Return articles dated between . . ." or author/research team ("Return articles authored by . . .") – both of which can serve us well in our search (see Final Considerations for Replication, pp 24–26).

Google search screenshot

[BOOK] **Bilingualism** and Minority-Language Children. Language and Literacy Series.
J Cummins - 1981 - ERIC
ABSTRACT This handbook provides an introduction to research findings related to bilingualism in minority-language children, and describes the implications of these findings for issues of current concern in Canadian education. Bilingualism is defined as the production and/or comprehension of two languages by the same individual. The phrase" minority-language children" refers to children whose first languae is different from the ...
☆ 99 Cited by 602 Related articles

[BOOK] **Bilingualism** and the Latin language
JN Adams - 2003 - books.google.com
Since the 1980s, bilingualism has become one of the main themes of sociolinguistics-but there are as yet few large-scale treatments of the subject specific to the ancient world. This book is the first work to deal systematically with bilingualism during a period of antiquity (the Roman period, down to about the fourth century AD) in the light of sociolinguistic discussions of bilingual issues. The general theme of the work is the nature of the contact ...
☆ 99 Cited by 780 Related articles All 3 versions

[BOOK] Language mixing in infant **bilingualism**: A sociolinguistic perspective
E Lanza - 2004 - books.google.com
This book addresses the issue of language contact in the context of child language acquisition. Elizabeth Lanza examines in detail the simultaneous acquisition of Norwegian and English by two first-born children in families living in Norway in which the mother is American and the father Norwegian. She connects psycholinguistic arguments with sociolinguistic evidence, adding a much-needed dimension of real language-use in ...
☆ 99 Cited by 568 Related articles All 3 versions »

Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems
E Bialystok, G Luk, E Kwan - Scientific studies of reading, 2005 - Taylor & Francis
Four groups of children in first grade were compared on early literacy tasks. Children in three of the groups were bilingual, each group representing a different combination of language and writing system, and children in the fourth group were monolingual speakers of
☆ 99 Cited by 478 Related articles All 9 versions

Phonological acquisition in Malta: A **bilingual** language learning context
H Grech, B Dodd - International Journal of **Bilingualism**, 2008 - journals.sagepub.com
... monolingual controls, their error rates being similar to those reported for other studies of **bilingual** children ... Further, the sequential **bilinguals** made more consonant errors than the simultaneous **bilinguals** and at least one ... 158 INTERNATIONAL JOURNAL OF **BILINGUALISM** 12 (3 ...
☆ 99 Cited by 49 Related articles All 7 versions

Effects of speech practice on fast mapping in monolingual and **bilingual** speakers
PF Kan, N Sadagopan, L Janich, M Andrade - Journal of Speech, Language ..., 2014 - ASHA
Purpose This study examines the effects of the levels of speech practice on fast mapping in monolingual and bilingual speakers. Method Participants were 30 English-speaking monolingual and 30 Spanish–English bilingual young adults. Each participant was
☆ 99 Cited by 12 Related articles All 19 versions

FIGURE 2.1 Google Scholar screenshot: bilingualism.

In our example on bilingualism, we can already begin filtering out some of the papers that are most clearly not experimental research: those listed as "books" may be compilations of papers (perhaps including experimental research) or perhaps state-of-the-art reviews. However, you may well need to click through the examples given to be sure about the nature of the content. The Cummins (1981) book appears to be a (doubtless by now out-of-date) state-of-the-art review of research which may have limited interest for our present needs. Adams (2003) would seem to be an historical treatment of the bilingual use of Latin. Clicking through the Lanza (2004) entry, however, takes us to an interesting case study of two children acquiring Norwegian and English. Bialystok, Luk, & Kwan (2005) catches one's eye. It is worth noting (and clicking on) the number of citations to see how far the paper is significant enough to have been – and continues to be – cited (see Final Considerations for Replication). While not a definitive sign of importance, the citation statistics may well be indicative of a study that remains of interest to the field and perhaps continues to have an influence upon it. In the case of the Bialystok *et al.* paper, we see a fairly recent paper that has been cited 478 times and (clicking through) continues to be cited up to the present in books and papers which themselves are widely cited. We might begin to think this paper has the kind of continued significance which might make it at least worth reading in the way we describe below with a view to sounding out its needs and potential for replication. Equally at this point we would want to note exactly *where* it has been cited because these same papers may also give us clues as to these authors' opinions on the paper. Finally, again, our reading of the paper should take in the considerations mentioned below (Final Considerations for Replication).

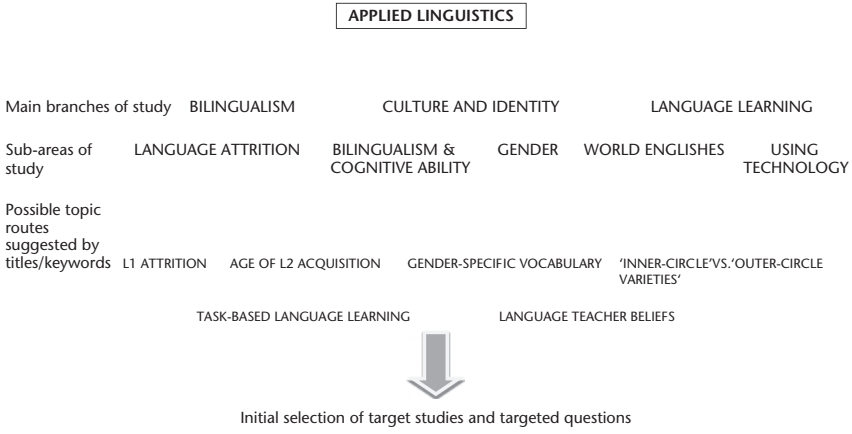


FIGURE 2.2 Sample routes to selection of your study.

» Activity 4: GOOGLE SCHOLAR SEARCH

Choose ONE of the following sample AL fields of study and use the term to carry out a Google Scholar search to establish one sub-field which interests you. *Then* use the search outcome to narrow down your search to two or three possible target studies.

- (Applied linguistics and) language assessment;
- (Applied linguistics and) L1/L2 interference;
- (Applied linguistics and) L2 pedagogy;
- (Applied linguistics and) intercultural issues;
- (Applied linguistics and) technology.

State-of-the-Art Critical Reviews

A further route is to generate a list of potential articles for replication from recent state-of-the-art critical reviews of the literature. The best sources for such reviews are journals (such as *Language Teaching* or *Annual Review of Applied Linguistics*) which encourage the author to appraise the research selected rather than merely list them in abstract form. They can usually be expected to highlight the most significant contributions to the field as well as indicate where research is lacking and – of particular interest in our search – what aspects of the studies remain to be further examined. It is worth learning how to spot these critical – but sometimes implicit – observations for they may well indicate what remains to be researched and, crucially for our present needs, the desirability of a replication study or clues as to what other work needs to be revisited.

» Activity 5: USING STATE-OF-THE-ART REVIEWS

***Read* pages 312 through 315 from the state-of-the art review of research on young L2 learners in Asia by Butler (2014).[1] *Underline* as in the example below what the author suggests are possible limitations and/or potential modifications to a subsequent study.**

Chow, McBride-Chang, & Burgess (2005)[2] conducted a 9-month study of Hong Kong kindergarteners in order to examine the relationships among three phonological processing skills in Chinese (i.e., phonological awareness (PA), rapid automatized naming, and verbal short-term memory) and word reading in Chinese (L1) and English (L2/FL). After controlling for visual skills that were considered to be an important predictor for Chinese reading, the authors found that the three phonological processing skills were moderately

associated with word reading in both Chinese and English, and that the association remained stable over time. In addition, PA in Chinese (measured by a syllable deletion test) was a relatively strong predictor for word reading both in Chinese and English, suggesting that PA is important not only for learning to read in alphabetic languages but also for reading in Chinese (a morphosyllabic/logographic language). However, the authors acknowledged that only syllable awareness was tested in this study. In Chinese, a syllable constitutes a morpheme that carries the semantic information of a word. Reading was also restricted to only word-level reading in their study.

Customized Calls for Replication and Published Replica Studies

A second source of critical review are papers in which an author specifically targets an area or group of studies for replication. As replication research is a somewhat novel methodological approach in AL, such dedicated position papers are rare. However, the journal *Language Teaching* (Cambridge University Press) runs a regular "replication research" feature where authors recommend certain studies for replication and discuss why there is a perceived need (e.g., Leow 2015 (volume 48.1); Bitchener & Knoch 2015 (volume 48.3); Gass & Valmori 2015 (volume 48.4); Ferris 2015 (volume 48.4)). These papers make for interesting reading as the instructions to authors require them to choose:

> a study or studies for replication which is a significant contribution to the field, and so needful of replication, in terms of its content and/or its impact on the field, has been published in a refereed-journal within the field and can be readily accessed for reference.
>
> *(Instructions to Contributors)*

These papers have the considerable advantage that they can not only be used as a potential source of a replication study, but also as a model for setting out the need for a replication in your eventual writing up of your work prior to submission to a journal (See Chapter 6). You will find an initial section in each paper which explains where the original study (and any other replications) fits into current knowledge in the field. The next section tells us the main details of that original study to be replicated in terms of how it approached the problem and framed the research question(s). Enough information is provided here to give you an idea of what went on, why, and what the outcomes were. If you are then interested in proceeding, you will need to read that study in detail (see Final Considerations for Replication). References are then provided to where it can be accessed in the references list. Authors then go on to tell the reader how they suggest the replication might best be framed, pointing you in the direction of

what kind of approach(es) might bear useful (and publishable!) replication fruit, and what elements of the original study might be usefully varied.

Published replication studies will be a useful source of potential recommendations. Authors will typically have reviewed the background literature to their own replication, pointing out the gaps in the literature and eventually focusing on their questions arising from the original study. Along the way they will often have provided similar indications of the need for replication in other studies. Equally, in discussing and concluding a paper, you may well read the author suggestions about how additional replications might further refine what has been revealed so far. In the replication studies we will refer to in the text, we will meet the typical "Review of the literature" section and, on the way to discussing the target study, the authors would then endeavor to highlight unresolved aspects of closely related work as well. Likewise, at the end of a replication, you would expect to see the typical "Limitations" section, where you might even find the authors suggesting further replications with subtle variation changes which they feel might contribute even more useful data about the target study replicated. Chapters 6 and 7 discuss these aspects in considerable detail when we take you through the steps of executing and writing up a replication study.

» Activity 6: ASPECTS TO CONSIDER ABOUT YOUR SELECTED STUDY

1. **TICK which of these aspects of an experimental study might make it useful for replication? Explain why you think so, using the discussion questions provided.**
2. **Then CHOOSE SIX characteristics you think are the most important and justify your choice.**

- A. The author(s) of the original study are considered key researchers in that field.

Think about the extent to which this suggests a replication would be needed or useful.

- B. The general topic of the original paper is one that continues to generate much debate.

How far is the topic – rather than the study – going to be of interest in terms of replication?

- C. The paper itself is cited in work around the same period.

Is the paper STILL being cited? What difference might this make to your decision?

- D. The original paper's findings are not consistent with previous or subsequent work in the area.

Why might this be interesting grounds for considering replication?

- E. The original paper continues to be cited in publications.

Think about why a paper such as this might continue to be of enough interest to be regularly cited. Is the criticism of the paper all in one direction only? What conclusions might you draw from this "popularity" as regards the possible usefulness of revisiting it?

- F. The references at the end of the original paper are now out of date.

How important is this observation with respect to the need for replication? Are there any circumstances when this observation might signal a possible need for replication?

- G. The original study is cited as one of the most significant examples in practice of a particular theory.

How far might the continuing relevance of, or interest in, the theory direct our thinking as regards replication?

- H. The original study identified limitations.

How might these limitations set us thinking about replication needs?

- I. The journal in which the paper was first published is a prestigious one.

Why might the reputation of the journal affect our decision? Think about the impact and potential audience involved.

- J. The participants chosen/assigned in the original study were very similar to those I work with.

How similar would they need to be to justify a replication on this basis? Is this reason enough to proceed to a replication?

- K. The statistical analysis used on the original data can now be improved upon.

How might a different available analysis or procedure signal the potential usefulness of a replication?

- L. A small number of participants in the original study.

Think about how the number of participants might affect the outcome of a study and the generalizability of the conclusions.

- M. The original study targeted students learning ESL (English as a second language).

How might replication in other language learning contexts or L2 prove interesting?

- N. The study reports that results were not statistically significant.

How might the apparently unsuccessful outcome be interesting with respect to further replication attempts?

- O. It would be interesting to use the methodology in my local language learning context to see what happens.

Would this be replicating the study in the terms we have described above in section 2.1?

- P. I wonder whether an intervening variable, such as participants' educational background, might have affected the outcome of the original study.

(continued)

(continued)

Which specific variables do you think might have affected the result, and how might a replication attempt to discover this?

- Q. I wonder if adding a further source of data would provide additional interest from a replication study?

What other data sources do you think might be added to the original study and which might conceivably provide us with useful data?

- R. Effect size data is not presented or is not convincing.

How might a replication which provides such a statistic add to our knowledge?

Final Considerations for Replication

As we have read in the Introduction, *any* experimental research can – in theory – be replicated in some form or another. It is quite another question whether it *needs* to be, and what further useful data can be gained by doing so. Aside from their perceived merit, there are other considerations or questions arising out of the original study's design and publication which we need to be aware of when selecting a candidate for replication.

Firstl, we need to **appraise the continued relevance of the research in question**. To do this, we need to have read generally around the subject and, specifically, the objectives of the study in question. Not every published research study has sufficient continued scientific significance to merit replication. Conversely, if we find the current literature still cites and/or refers to that study when discussing the specific area in question, we may deduce it is still influencing thought and research practice and stimulating research. Studies whose aim was to extend the validity or generalizability of a particular finding might be good places to begin, as are those that test interventions that have a potential impact on learning outcomes.

Publication dates will also likely be of importance here – albeit not a decisive criterion for including or discarding a study. Bear in mind as you carry out your search that the nature of publishing often means that a paper published this year or the year before may actually be reporting on research carried out three or even four years before that! The further back your search goes, the more chance that the target research has already been superseded by later events or more recent research. This is not to say that older, seminal papers that have not seen any replication should therefore not be addressed. They are worthy of attention, particularly if they are still cited and appear to be of continued importance for the field.

A caveat here is the question of **access to the original data and data collection materials**. As we will discuss later, this can be one of the most difficult steps in producing a replication study. It figures that if we want to replicate a study accurately,

we need to know exactly how it was carried out, from participants/data selection, through instruments, procedures to results, and analysis (see also Chapter 6).

It is a fair assumption that the further back you delve in the published archives, the more difficult it will be to obtain this kind of detailed data and procedures from the original author(s). There are severe space restrictions on most printed journals these days and authors may simply not have the space to present all the detailed data you need: for example, perhaps we can read about the number of participants but their learning context/history may need more detail. If the information is not available as supplementary data on the publisher's website, in online repositories such as IRIS (www.iris-database.org), or on research project websites such as LANGSNAP (http://langsnap.soton.ac.uk), you will need to be able to contact the original authors and request it.

Be on the lookout, too, for studies where **an unexpected or unusual outcome** is reported. Such outcomes from well-designed and rigorous studies are interesting in the sense that things appear not to have come out the way the authors were expecting or the way the literature suggested they might; therefore, it is likely further examination of the question is warranted and to be welcomed and would lead us into critiquing some aspects of the study with a view to its future replication (see section in Chapter 3, Routes to Replication: Reading with Targeted Questions). From such unexpected outcomes (perhaps a lesser effect than anticipated or the influence of an unforeseen variable), there might well be a significant discovery made which is a game-changer. We can only begin to get to a more confident conclusion, however, by undertaking suitable replication studies of the effect or intervention which might also average out any biases or errors in the original.

Of course, we might be equally attracted to studies where we feel exaggerated claims for a cause–effect or a particular intervention are made. Much experimental research we read about shows a positive outcome, null hypotheses rejected, or interventions shown to result in gains in learning. There is evidence that the "file drawer problem" observed in social science research overall, which is a tendency to publish positive results rather than negative or non-confirmatory outcomes, is just as prevalent in AL research output. Research that presents negative or unexpected outcomes is often seen to be of less interest and, therefore, may not even reach the published page (see Chapter 8). It is worth noting that moves in journals such as *Language Learning* to initiate (as from 2018) a Registered Reports section may alleviate this:

> in which a substantial part of the manuscript (including the methods and proposed analyses) is pre-registered and reviewed prior to the research being conducted. This format is designed to reduce bias and other questionable research practices, particularly in deductive science, while also allowing researchers the flexibility to conduct subsequent unregistered analyses and to report serendipitous findings.

But in negative or unexpected outcomes there lies an opportunity for the researcher interested in replication. In much scientific experimentation we should expect our experiments to "fail" a lot of the time; it is important to find out why before moving on to something new. In our search for a worthwhile replication, the negative, unexpected result may often take us down those unpredicted but fruitful paths where the most interesting discoveries are made. A replication study based on such a paper can only help enlighten us.

You will obviously want your replication study to have the same or greater methodological and analytical rigor as that of the original. It would be sensible to select a paper which uses **methods and/or statistical analyses with which you are familiar** and that still holds currency which you can, in turn, manage in your own replication – or obtain external help to execute. You will need to analyze the statistical complexity involved in carrying out the original study both to understand how the results were arrived at and be able to proceed to your own subsequent analyses. You will also need to select based on the particular research design or methodology used in the original study. Thus, the feasibility of your carrying out a descriptive, quasi-experimental, or experimental study in your context will need to be calculated as well as the practicability of conducting interviews, questionnaires, surveys etc.

The **medium or source of publication** is a further factor to be taken into account when we select our target study (see Chapter 3). Scholars would usually aim to publish their work where it is likely to get the maximum attention and dissemination. A "quality" journal is more likely to have an editorial board and group of reviewers or referees who the editor would look to to evaluate the strengths of the claims made on the original study. Searching for a paper in a quality journal is important, therefore, as it would more likely turn up a paper with sufficient rigor and internal validity to encourage us to build on an already sound piece of work. Such a paper might also be a stable platform from which to launch a replication which aimed to address some of the limitations expressed in the paper and strengthen (or temper) some of the generalizability claims found therein.

Notes

1 Butler, Y.G. (2015). English language education among young learners in East Asia: A review of current research (2004–2014). *Language Teaching*, 48.03, 303–342.

2 Chow, B.W.-Y., McBride-Chang, C., & Burgess, S. Phonological processing skills and early reading abilities in Hong Kong Chinese kindergarteners learning to read English as a second language. *Journal of Educational Psychology*, 97.1, Feb. 2005, 81–87.

3

PLANNING YOUR REPLICATION RESEARCH PROJECT

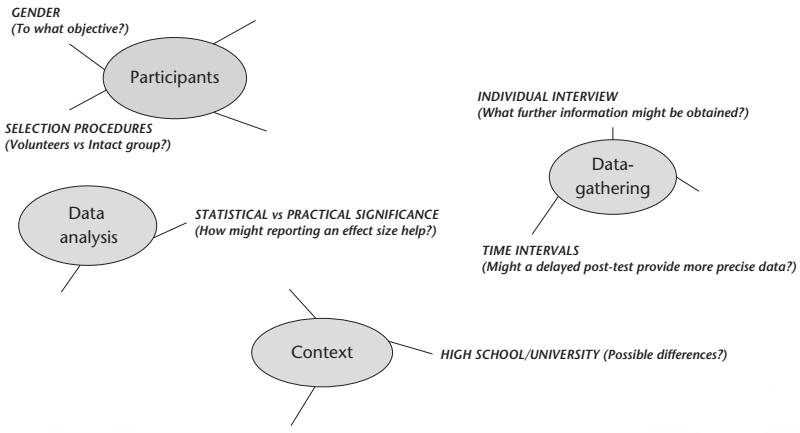
3.1 Unpacking a Study Selected for Replication

At this stage we will assume you have followed the steps in the previous chapter and are now ready to focus in on a "short list" of two or three studies that have passed muster in your initial search and selection. We would now have the studies in front of us and be ready to address critically with a view to choosing one for replication.

Before we do, however, let us briefly brainstorm what aspects of an experimental study might be susceptible to replication, and why.

» Activity 7: WHAT VARIABLES MIGHT BE USEFULLY MODIFIED IN A REPLICATION STUDY?

Here is a "mind map" starting with four basic aspects of an experimental research study. Brainstorm features of each and then form questions such as those in the examples to establish possible routes for replication.



3.2 Routes to Replication: Reading for Awareness-Raising

Our approach to this, more focused, stage of critique will be to address a number of aspects in the study through a close reading of the text. As an initial recommended response strategy, however, and before embarking on any specific questions aimed at establishing the potential for replication (see Routes to Replication: Reading with Targeted Questions), you are encouraged to take notes on either side of the text and record your spontaneous reactions to what you read *as* you read it – much the same way as if you were engaged in live dialogue with the author/researcher.

One of the central aspects of this approach to eventual study selection, therefore, is for you to stimulate your critical faculties through such a stream of consciousness on your first close encounter with the text. The idea is that these collected thoughts, questions, and observations flagged up throughout the text – together with the more targeted questions that follow – then form the basis for our decision as to what aspects might reward closer observation, and whether the whole study might then be flagged up for some kind of replication.

This initial awareness-raising reading strategy is designed to make you become familiar with the text and train you to be constantly responsive to what you read. This meticulous approach to reading a paper is needed if we are to pause, assimilate what we have just read, and then form a response to it that might help us decide what might form the key aspects of any attempted replication.

In the example in Figure 3.1, we show how this works with an abstract from one of the key studies we referred you to in the Introduction.[1] We would expect the abstract to serve as an initial indicator as to whether the study is sufficiently relevant to our current interests in replicating the study and – if so – to enable us to make written notes on the page about what aspects/details of the study we would need to watch out for as we read the main body of the paper with a view to possible replication.

Our initial spontaneous reactions to the text are noted down in the column space on either side of the text.

3.2.1 Routes to Replication: Reading with Targeted Questions

Our preliminary reading should have now identified a number of aspects of the target study that have attracted some initial responses.

Here is a summary of the Carter, Ferzli, and Wiebe (2004) study

> Carter, Ferzli and Wiebe (2004), published in the highly-respected journal *Research in the Teaching of English*, is a study that has been often cited in the literature as one of the few that have researched the success (or not) of genre-based instruction. This need for genre-based knowledge in academic English production remains of great interest to the field but this specific

FIGURE 3.1 Reading a paper: awareness-raising.

paper, by being often cited as one of the first to attempt formal online instruction of the same, may well reward closer attention and possible replication. In the present case, the authors have concluded their paper with an interesting challenge. They wonder whether the treatment would be as successful with "English L2 students, the primary focus of English for Specific Purposes".

The quasi-experimental study investigated the effect of genre-based instruction on the writing of university students learning disciplinary-specific writing. One aspect that makes it an initially attractive candidate for replication is that it comes with enough detail regarding method to permit replication. Two groups were set up, one receiving a handout on preparing lab reports and specially designed instruction online software (*LabWrite*) and the other (control) group receiving only the handout. Sample lab reports

> from both groups were analyzed and a final questionnaire administered to obtain individual attitudes towards writing lab reports. Results showed the treatment group producing reports judged to be of better quality; the survey showed their attitudes towards the writing to be more positive than the control group.

We now take you through a closer, critical address of this paper, this time explicitly to isolate and question a select number of features and variables that make up the total research situation in the target study, and that potentially affect the outcomes of the research and which, therefore, might warrant consideration in a potential replication study.

Even good research designs cannot account for *all* of the many threats to internal validity and just one of these can have impacted outcomes sufficiently to warrant adjustment in a subsequent replication. On the other hand, there is the risk when carrying out a replication study that you may be repeating a study that is so poorly designed or executed anyway that it would not merit the effort. Again, your close reading of the text will help you to choose wisely!

All this means that, as we ask these questions with an eye to the usefulness of a potential replication study, we are also indulging in research critique. The more questions we can see answered in the study as it is described, the greater apparent soundness that study would seem to have with a view to our replicating it. Rather more common, however, would be for us to find unanswered questions arising from our reading which require more details to be obtained – perhaps by contacting the original authors or accessing a web repository or project website. Nevertheless, if you find yourself questioning most of the procedures, methods, and analytical process, this may be a sign of a study which currently lacks sufficient detail to be considered further as a possible replication.

The final practice activity in this chapter (pp. 41–46) will see you apply these targeted questions to a complete paper. This concluding stage in selection should help focus our thinking down to what a replication of the paper might look like and what this might reveal as a result of modifying the factor or variable in question.

3.2.1.1 Participant Characteristics (p.400–401)

Variables including age, gender, L1, previous experience (language learning and other), personality, motivation, and socioeconomic background would all be expected to vary greatly across participants and studies. As we read our studies, we might think about how far these variables could have affected outcomes in the study in question. For example, on page 401 in the Carter *et al.* study, we read tantalizingly few details about the participants: we know they are "university students", presumably first year, probably all around the same age, but nothing else is known about them. Such information might have been deemed unnecessary as it might have had no effect on an outcome which showed the

intervention treatment to have worked with the experimental group. It might, for example, be interesting to see how far previous experience, in the shape of students' knowledge of the genre prior to instruction, might influence the outcome. You might then annotate that section of the text with the targeted question(s) arising (e.g., "*I wonder if a pre-test of previous genre knowledge would provide some more useful information on the true extent of any treatment success?*"). A further question for a replication might be whether the LabWrite treatment would be as successful with participants using a similar genre but of another sciences major (i.e., other than biology). The authors make no claims for the efficacy of the treatment with L2 students, and it would clearly be interesting to see how useful it would turn out to be with this audience. In such circumstances, we might assume the LabWrite treatment might need simplifying or adjusting to the L2 proficiency level of the students to meet the needs of such students in some way. An adapted form of LabWrite tested in a similar way could be an interesting replication.

3.2.1.2 Sample Size

Here we might be interested as to whether the sample is large enough for us to have confidence in the outcomes reported. Large samples – involving large amounts of participants perhaps in different locations – are expensive to set up and often take longer to carry out. Smaller samples have the advantage of offering a closer view of the participants together and the disadvantage that they are more likely to result in statistically significant results that can represent false positives (indicating that a given condition or outcome is present when it actually is not). The reason for this is that each member in a small group will perforce have a larger effect on the general group performance than would have been the case in a larger group. Such outcomes in smaller sample studies need to be treated with circumspection. In the case of greater numbers of participants, you might want to consider what data about individual reactions might *not* normally be obtained without additional data-gathering such as interviews or think-aloud protocols.

The number of participants we read about in a study will tell us something important about the study's "power" and/or generalizability and whether the outcomes might be further strengthened by attending to this aspect in any future replication (see p. 55 below).

In the present case, we read that numbers were reduced to "approximately 80 students . . . to limit possible teacher variable" (sic). While it is not clear how far such a variable might have affected results, the numbers involved would in principle appear to give us confidence in the outcomes presented.

Although we might assume all the students in Biology 181 or 183 were of similar abilities, and writing skills, no pre-test or other information confirms this. A replication with L2 students would also want at least to ensure similar proficiency levels, independently measured, ideally. Furthermore, a replication might want to match participants between the two groups for key characteristics such as

gender or L2 proficiency level. If randomized allocation to control and treatment groups is not possible, matching participants will go some way to permitting increased comparability between the two groups. Thus, data collection procedures can be varied if we consider them potentially limiting.

3.2.1.3 Research Histories of Participants

A common problem with all participant selection is establishing the extent to which any prior treatment or classwork may affect performance in the present study. Similarly, in this targeted question you would be interested to know whether outcomes might have been affected by the voluntary or other nature of participation in the study. Thus, for example, the motivation of participants who have been offered some kind of reward for their involvement might be expected to differ from volunteers. We might also assume that participants in certain academic environments were likely doing "other things" elsewhere with their L2 during the intervention (particularly in ESL environments) which might have affected outcomes and which we could control for in a replication. Ideally, of course, you would also have wanted the experimental and control groups to have experienced the same or similar history factors, including teacher/researcher effects and classroom/laboratory conditions. Care was taken in the present case to avoid cross-contamination by giving the control group sessions in the previous semester to the treatment group. The conditions for the treatment and the control seemed rather different, however, as the authors themselves admit. The control and treatment groups were taught by the same professors/instructors but the former were left to consult the guidelines for writing up their reports on their own and asked to "refer to the handout during the semester when questions of how to write reports were raised", while the treatment group accessed a web-based course with more dedicated, explicit, personalized instruction. One might suggest the latter could be seen as a more motivating, as well as more guided and interactive, experience and one that the participants might therefore have been more interested in consulting regularly.

In a replication we might want to even the odds more: perhaps the control group might be offered more internet-based interaction with their instructors – rather than left to their own devices – with the latter providing comments on their work during the semester. It also seems that the composition of the two groups might be improved upon in any replication as the control group turned out to have a greater range of "differences in . . . curriculum" than the treatment group.

3.2.1.4 Length of Treatment/Intervention

Longitudinal studies are the exception rather than the rule in much of the research we read about. There are sound reasons for this: they take longer and might therefore take up more classroom time away from the academic participants. They also

require closer control and/or monitoring as they become more susceptible to the kind of history factors mentioned above. Conversely, you might address the fact that a treatment or intervention needs time to obtain definitive results – successful or otherwise. In this sense we would be interested in how far the research design included suitable post-test measures or events (see page 35).

The study tells us only that the treatment continued in the spring semester without specifying the amount of hours or frequency of classes received by either group. A replication would probably want to specify and perhaps test this time period as a new variable. Elsewhere we are told that the control group was "introduced to writing lab reports on the first day of the lab class" but there appears to be little or no supervised/taught revision beyond this day. The treatment group, we understand, received this "integrated into" the course, which may indicate rather more constant, and guided, reference to the materials. This difference in access and use of the materials offered may also have affected outcomes.

3.2.1.5 Group/Physical Setting

Many variables are at work here. These can range from the way participants were selected for the different groups, the room, or setting for the intervention (would, say, a familiar classroom be expected to present the same outcomes as the researcher's office?) to the composition of each group (e.g., how did the control/experimental groups differ, was the sample selected representative of the target population for the study, and was the study adequately "blinded"?). The study under consideration does not give many details as regards selection; we seem to be seeing intact classes here ("students registered for Biology 183 . . .") from which "the study focused on 80 students from 4 lab sections". How the selection was made, and what the size of the original cohort was, is not clear, and before a replication study was undertaken we might want to try to extract more information on this from the authors.

3.2.1.6 Control Agent

Many of the studies we read will involve work carried out in classroom contexts. A control agent will be the person who effectively runs the process in that context – it might be a familiar teacher or a complete stranger to the participants. The situation can be further complicated if either person is the researcher. "Complicated" in the sense that the researcher might be inclined to "see" only the positive in what he or she sees or in the data obtained. We could suggest, however, that some treatment variable outcomes might be further enhanced if the person giving the treatment is the person who designed the intervention. Conversely, a replication which did *not* involve the researcher in any direct role might achieve greater internal validity since the results can then be interpreted as independent of any apparent predisposition to a positive outcome.

In the case which concerns us, we read that care was clearly taken to make sure the same professors taught both groups, "Control and treatment students were taught by the same professor in the lecture sections and the same 2 instructors in the labs (each instructor having 2 sections), with identical course syllabi, labs, and assigned lab reports", and the setting would seem to have been familiar – the normal classroom or lab participants were used to. The authors themselves, however, make the point in their Discussion/Limitations section that "the in-class instructors were 'insiders' of the labs" but insist that this would not have affected results as it was the website which provided the instruction throughout. A replication of the study might want to see how far similar results are obtained with non-specialist instructors (particularly in an ESP (English for specific purposes) or EAP (English for academic purposes) situation) even if these are assigned an apparently secondary role which is mainly "to encourage the use of the website through assignments".

3.2.1.7 Specific Task Variables

A number of variables are involved in the material circumstances surrounding the course of the study. Depending on the objectives pursued in the intervention one might think aspects such as the format for presenting information (paper/online) as well as the conditions of that task (e.g., against the clock/assessed grade to be awarded which contributes to year grade) may impact outcomes. In the current paper the treatment was given from a website which participants were encouraged to consult regularly. Compared to the kind of autonomous learning approach for the control group which:

> . . . was typical of college lab classes. Students were introduced to writing lab reports on the first day of the lab class with a one-and-a-half-page handout listing the sections of the report and brief descriptions of each section . . . follow the handout as a guide to writing their reports, referring to the handout during the semester"

we might suggest the treatment group received a novel, cohesive, and far more continuous kind of instruction ("integrated into the laboratory activities of the Spring semester course . . . LabWrite Post-Lab also gives students access to a step-by-step guide for writing each section of the report."). Using an L2 cohort, a further test of the success of the treatment might be made by having the treatment and control groups access any guidance through the same medium.

3.2.1.8 Instructions and Cues

Often omitted in detail in papers, the instructions received by participants – and the way these might be interpreted – can prove crucial to outcomes. Deliberate or

unintentional variation in these instructions would amount to manipulation of an independent variable. These stimuli would preferably be received uniformly by all the participants receiving or not receiving the treatment but, in practice, this level of uniformity is difficult to achieve given the individual interpretations which might be available. As we have seen, the "physical setting" can also interact with the instructions here: for example, instructions given by means such as listening cues may not be uniform because of the nature of the medium itself and the existing listening conditions in the room. The instructions themselves should also be presented in a way that we can be sure accords with the age and L2 (if these are given in the L2) of the participants. In this paper, the treatment instructions seem to be far more guided and detailed than those given the control group (p. 401–402). The control group had to rely on their own motivation and judgment to refer back to the handout; the treatment group appears to have been both oral and written while the control group seems to have been mainly reading. A replication might want to have at least the cues made the same for both groups.

3.2.1.9 Instruments/Tools Used (What Advantages or Limitations Are Admitted or Observed?)

As we read above, our hope is that enough information about the procedures – and in particular any data-gathering instrument(s) – is provided in the paper itself or accessible elsewhere. In our critique of this aspect – and more specifically with a view to discovering a need for replication – we will inevitably be working within the constraints on space imposed by the publishing medium. While it might be impractical to see a detailed sample of the materials used, it would be helpful to know where any materials can be accessed and to read an informative summary of the design development, piloting, items, raters and their training, scales, and scoring. In a replication, for example, we might feel participant responses could have been affected by the content of some of the questions or the speed with which they might have been delivered and want to redesign some, alter the order, or even the language in which they are presented. As with the instructions or cues (3.2.1.8), we might also feel that the technology employed in gathering the data could impact outcomes: you will come across many different vehicles for administering treatments, all of which may have affected the way data is presented and received visually and/or aurally.

3.2.1.10 Measurement

Outcomes are likely to be affected by the way individual or group responses are scored or assessed. We need to read about how responses were then weighted as scores to fully understand their significance on the results. This can help us decide how far, say, a Likert scale response over-limited the range of responses

and therefore might require tweaking in any subsequent replication. Similarly, we would be interested in reading when measurements were made, pre-, during, and post-treatment. Many interventions you read about are designed to test the effectiveness of a specially designed set of materials. As we saw, comparatively few longitudinal studies are carried out when such treatments are tested since the typical demands on their time available for university or college participants do not allow such lengthy interventions. However, when a treatment is being administered, it is often of more interest to see not only how successful or otherwise this turned out to be at various points during the intervention. Typically, we might be presented with a pre- post-test design in which only these two measurements are taken. While two measurements may show us an indication of improvement after a treatment, it will not show us if that improvement lasted beyond the intervention period and – crucially - was really "learnt" or its effect remained.

Our confidence in the data we collect is informed in part by its "reliability" statistic. There are two main types of reliability. The first, instrument reliability, refers to the extent to which a particular instrument is internally consistent. If a questionnaire is designed to measure motivation, for example, the items in that scale should generally "agree" with each other. The most common index used to measure internal consistency is Cronbach's alpha, which is expressed on a scale of 0 to 1. The other main type of reliability, inter-rater reliability, provides an indication of agreement among raters. We often find inter-rater reliability estimates when more subjective scales are used such as in writing or pronunciation assessment. Inter-rater reliability can be expressed as a correlation when working with scores; in the case of categorical rating schemes, we usually find percentage agreement or an index called Cohen's kappa, which accounts for chance agreement among raters. Values for kappa also fall between 0 and 1; for a guide to interpreting reliability estimates in L2 research, see Plonsky and Derrick (2016).[2]

If no such statistics are available, we can assess the instrument informally to decide whether we think the instrument is providing a consistent and accurate result. Does the instrument seem to promise accurate data-gathering from the way it is constructed? Do we think the interview questions show bias and/or do we think they can be answered accurately by the participants and scored accurately by the raters? Is the rubric or instructions, if used, sufficiently clear and able to be applied? Was the instrument piloted or field-tested beforehand?

In another paper by Bitchener and Knoch (2010)[3] the data instruments and details of measurement are presented in considerable detail in the appendix and the text itself, all of which facilitates any further replication. Attention has clearly been paid to the training of the raters in charge of marking the lab report output of both groups and satisfactory rater reliability statistics are presented. The results claim that the treatment group outperformed the control group on the holistic analysis of the lab reports and "learned the scientific concepts". However, we are dealing with a post-test-only research design. Apart from the incorporation of

some kind of adequate pre-test (see 3.2.1.1, *Participant characteristics*), a replication which incorporated more points of data collection and measurement might prove enlightening, as would delayed post-tests later in the year, or even in subsequent years – particularly considering we read that the participants were only in their *first* year at university. Increasing data collection points will also help the replication reveal any underlying evidence of growth trends in either group or change after the treatment.

3.2.1.11 Statistical Significance and Effect Size

As regards the analysis and results of the data, our targeted questions for replication would ideally focus on three aspects: (a) how was the data analyzed and could any adjustments or alternate analyses – in addition to the original procedures – be carried out in a replication to arrive at more valid outcomes; (b) what do any quantitative measures of the strength of a phenomenon reveal both in relation to the current study and in relation to the study being replicated; and (c) are there any other potential explanations for the results obtained.

a) We have already seen how the number of participants can increase (or decrease) the power or relative generalizability of a study. Our first port of call, however, will be the statistical significance claimed for the study. A value typically used in social science studies is $p<0.05$. There are – at the level of general research critique – a number of problems with the so-called "Null Hypothesis Testing" approach which you should be aware of (Norris, 2015; Plonsky, 2015),[4] but in terms of building targeted questions to establish the usefulness of a replication, we need initially to pay some attention to this value in the context of the research design (see below, b). While it is less likely these days you will find a published study which fails in its attempt to claim statistical significance for its outcomes, the failure to do so, or the appreciation of a weaker effect or relationship than expected are results that should call our attention in the context of finding a suitable study to replicate. Arguably, they are of more potential replication interest than "successful" outcomes!

Much of what you read implicitly defines success in empirical research in our field in terms of a "statistically significant" result. This creates a situation whereby we either do not read about, or are inclined to view with lesser importance, non-statistically significant outcomes. However, these outcomes are just as interesting, of course, to show us where our ideas are apparently *not* working and of great interest to those seeking to search out a study which might benefit from replication. A negative or not statistically significant outcome is not a failure if we see it as showing the researcher's assumptions had not been correct on this occasion – because they might be shown to be so in a suitable replication which attempts to "correct" potential errors. After all, the scientific method encourages us to frame testable hypotheses and then try to falsify them to determine how likely they are to be accurate. So-called "false negatives" (a test result indicates that a condition

failed, while it actually was successful) do happen, and it is worth replicating even negative results, but that certainly does not mean they are of any less scientific worth than positive outcomes. Indeed, it is possible the original researcher made some mistake in the original work, and you might then be able to "correct" that mistake and go on to make entirely new discoveries.

b) While attention to statistical significance is an initial point of reference, you will also read studies which describe outcomes in terms of statistical significance alone – without actually describing the strength or magnitude of the result. One of the problems of "Null Hypothesis Significance Testing" mentioned above is that it encourages dichotomous thinking: either a result is statistically significant, or it isn't. But this is only half the story: a more informative and better contribution to progress of knowledge is made if we then discover how big the effect is or how strong the relationship is.

Initial point of reference apart, your focus in most instances needs to be on this effect size. Looking at *p* values alone will not be enough to suggest a suitable direction for possible replication. More insidiously, as an outcome in a replication study, it will lead us to misinterpret whether an original study was replicated. By focusing on the effect size, we can begin more correctly to think about the *degree* of replicability obtained.

An effect size measurement (e.g., Cohen's *d*), together with a confidence interval or margin of error, is particularly useful when we compare replications of a study as, for example, we would be able to summarize a group of experiments or replications that used the same independent and dependent variable and then directly compare the effects across the studies. These results can then serve as a starting point for additional replication or follow up studies that further describe the psychological processes that underlie an effect and/or tell us more about the limits of those conditions. Small effect sizes might be down to weaknesses in the way a study was set up or data collected or analysed: a replication study may address these apparent flaws. Should an effect size measurement not have been reported, there are a number of web-based calculators which can be used given the minimum descriptive statistics (e.g., www.uccs.edu/~lbecker/; www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html; or www.psychometrica.de/effect_size.html). A subsequent replication can be set up to further test the strength or weakness of that effect. Knowing the expected size of an effect is important when planning a replication aiming to further enhance the strength of the effect or relationship in the original study. An effect size calculated from a large sample, for example, is likely to be considerably more accurate than that from a small sample. An effect size's confidence intervals can also provide useful information about the reliability or robustness of an effect size (see Cumming & Calin-Jageman, 2017).[5]

A principal advantage of comparing effect sizes across original experiments and their replications is that the different size estimates from each study can then

be combined to give an overall best estimate of the effect size. A number of such experiments and resulting effect sizes can then be synthesized into a single effect size estimate in a "meta-analysis". Even more useful would be for us to examine, in a series of replications and original studies, any differences between those with large and small effect sizes and attempt to explain why there might have been such a difference reported. This process is known as "moderator analysis" (see Oswald & Plonsky, 2010, Plonsky & Oswald, 2015).[6]

In the present case, we are presented (on p. 406) with a treatment group which outperformed the control group ("significantly more effective") at a probability level of $p<.0001$). While inter-rater reliability is calculated and is acceptably high, this first hypothesis outcome has no reported effect size. Consequently, we only see part of the story here: an indication of effect from the treatment, but no information on the size of that effect. A subsequent replication should seek to remedy this for the reasons stated above. A number of such replications would have the advantage of allowing us to summarize a series of experiments all with the same independent variable (the LabWrite treatment) and directly compare effects across these studies, regardless of the numbers of participants involved – invaluable knowledge when we are testing the effectiveness of a potentially innovative intervention such as this.

c) Science advances by our asking questions and querying others' explanations. We assess the explanations given by examining the evidence put before us, comparing it, identifying weak reasoning, highlighting claims that seem to go against the evidence, and/or suggesting alternative explanations for similar outcomes.

A crucial ability in reading any study – but particularly one which we are sounding out for possible replication – is that of critically addressing the outcomes as concluded by the authors. One of the key aims of research is to search out and discover, but not to *prove*. Therefore, we can expect the author of the study to "examine, interpret and qualify the results – as well as draw inferences from them" (APA Publication Manual, p. 26)[7] in a particular way – but by definition, there will be others.

Furthermore, while we may agree with the results themselves, we may want to disagree with what those results *mean*. If a treatment is claimed to work, are further replications needed to confirm the extent of that finding? For example, a researcher may suggest results were not affected by an intervening variable identified post facto but you may feel there is evidence they were. A suitably designed replication study might be useful to sort the discrepancy out. The researcher might also indicate in a "Limitations" section what was perceived to have affected results, and through these thoughts we can map out a useful replication study. We might also feel the kind of generalization (see pages 40–41) made here would be useful for the field, but cannot be justified given such aspects of procedure as the participant selection, group size, or pre- and post-test measures. Again, if we felt the contribution of the study merited it, a

replication which attempted to fine-tune some or all of these aspects would be useful. Then again, you may find yourselves comparing these outcomes to other similar research in the field: does the study leave out some important aspect of the question that perhaps renders the conclusions wrong? If the outcome goes *against* what other similar studies have revealed, there may well be further need for clarification in a replication.

Apart from what has been noted above, our present study provides us with a number of other points of critical address as well as acknowledged limitations which could prove useful starting points for a replication. On p. 408 we read the authors' acceptance that "these results certainly do not rule out the possibility of effectively teaching genre in more traditional ways . . . we await sound control-group studies that investigate this possibility". One of the many replications that might answer this call could see LabWrite pitched against a more direct approach in a teacher-fronted classroom and one that offered a more continuous and course integrated control (rather than indirectly through a student-centered methodology and out-of-class reading).

On p. 409 we also read the authors' acceptance that the results cannot as yet be extended to L2 students – ". . . the primary focus of ESP". This would be a potentially important application of LabWrite, of course, and as such would be a useful replication if participants could be gathered in a similar context ". . . without preexisting tacit genre knowledge". Finally, as noted above, one of the weaknesses in this study is the fact that there were no delayed post-tests in which we could have witnessed the continuing effects of the treatment. A replication study might usefully address the limitation that "we can make no claims for transferability to subsequent lab classes" of the treatment by such additional post-treatment testing across subsequent months and perhaps into concurrent science subject courses.

3.2.1.12 Does the Generalizability of the Study Need to Be Further Tested?

One "successful" instance of a treatment has little generalizability beyond that one case. We cannot really be sure of what might happen outside the present context without turning to replication. If we want to see whether results hold across a range of populations, contents, treatments, and time, and then make suitable recommendations to educators, we need to turn to replication to achieve the cumulative impact required. Many of the studies you read will have been carried out using student participants, very often within the boundaries of the university or college setting in which the researchers found themselves. Such participants are relatively easy to recruit, but what one gains in convenience can also be lost in generalizing beyond this group. Student or school participants have a number of characteristics as L2 learners which differentiates them from those learning outside those institutions, not least in terms of proficiency level, instrumental motivation,

time devoted to learning and, of course, other L2s. So many variables are involved that the needful path of generalizing from a study will inevitably require a slower pace and detailed approach to replication – and a number of replications at that. As we saw in the previous chapter, with replications, we are best to talk of *degrees* of replication across studies; in such a slow accretion of results each replication is designed to provide a little more knowledge about the extent of effectiveness – and the generalizability thereof – found in the original work. As Plonsky (2012, p. 117) points out "For this reason, we must accept and even embrace a more incremental and . . . patience-testing pace towards cumulative knowledge".[8]

In the Carter *et al.* paper we have already mentioned the possibilities of replicating the study with L2 participants without previous genre knowledge, but the apparent usefulness of LabWrite (suitably adapted) shown here with students in biology labs needs to be further tested with other science subject classes.

» Activities 8a through 8h:
Key paper 1: J. Bitchener and U. Knoch (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, 19, 207–217.

Focus, first, on the "Abstract" and carry out the same initial awareness-raising reading strategy you saw in 3.2. Routes to Replication: Reading for Awareness-Raising (p.28) above using the space around the text for your annotations. Once you have completed your own reactions, compare these with our suggestions in Figure 3.2 overleaf. You are not looking for things that have been omitted or should be in the abstract itself , but rather things you intend to go on and find or think about when reading the body of the paper.

» Activity 8b: PRACTICE AWARENESS-RAISING: THE INTRODUCTION AND RESEARCH QUESTIONS

Focus now on the "Introduction" (Aims), "The study" (Aims), and the following Research questions/Hypotheses and carry out the same initial awareness-raising reading strategy you saw in 3.2 Routes to Replication: Reading for Awareness-Raising (p.28) using the space around the text for your annotations. Once you have done so, take a look below at some of our own thoughts as we were reading these sections.

(continued)

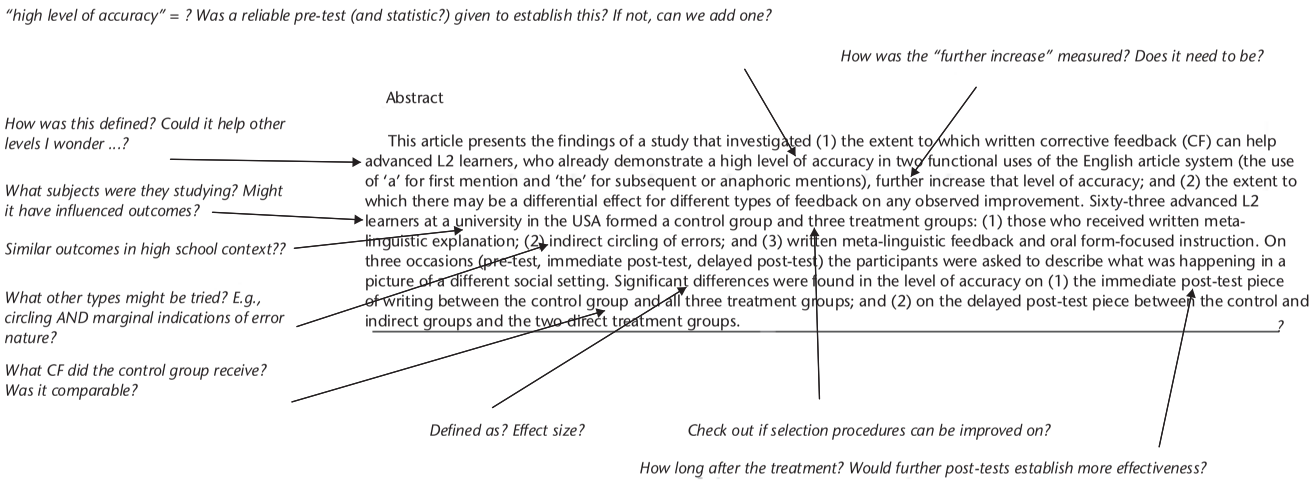


FIGURE 3.2 Practice awareness-raising: the abstract.

(continued)

- The authors talk of helping L2 writers "raise their level of performance" (208) through written CF (corrective feedback). There are several such phrased expressions to describe improvement in the text (e.g., "gain greater mastery" . . . "measured by improved accuracy"). I wonder if we might define this better in a replication perhaps via an established proficiency test (together with "lower proficiency" further into the text)?
- "It could be argued that advanced writers have a greater attentional capacity than intermediate writers . . ." (208). Are both these proficiency levels adequately defined in the rest of the paper?
- "The second aim of this study was to see if certain types of written CF are more effective in helping L2 writers improve the accuracy of their writing of new texts over time" (208). Does "over time" not assume a sustained benefit? How far is this demonstrated in delayed post-treatment testing?
- ". . . investigate the extent to which written CF can help advanced L2 writers . . ." I wonder if this might also extend to advanced L2 speaking, too?
- We read in one research question (211) "Does advanced learner accuracy in the use of two functions of the English article system improve over a 10 week period *as a result of* written CF?" (our italics). Can we be sure it was all down to this? Were other possible intervening variables accounted for in such a university environment? Could a replication control for some/any?
- The ten-week period is maybe a little short? Could we look for longer-term effects in a replication? Perhaps follow participants throughout the university year?

» Activity 8c: PRACTICE AWARENESS-RAISING: THE CONTENT AND PARTICIPANTS

Now focus on the section entitled "The study: Context and participants" and "Target structures" and carry out the same initial awareness-raising reading strategy you saw in 3.2. Routes to Replication: Reading for Awareness-Raising (p.28) above using the space around the text for your annotations. Then use the targeted question sections *Participant Characteristics* and *Sample Size* above to help guide your responses. Once you have completed your own reactions, compare your answers with our suggestions below.

- We read that the participants come from the ESL department of a large US university (211). Would EFL (English as a foreign language) students yield similar results? Other ESL participants in the US could be a useful target in a replication, as well as ESL, but in another country?

(continued)

(continued)

- The participants are claimed to have had "a high degree of motivation" (211). How reliably was this measured? Might this be improved upon in a replication? Would the treatment work equally with *less*-motivated participants?
- In "Target structures" we read about the choice of functions as being because: "L2 writers across English language proficiency levels (including those at an advanced level of proficiency who demonstrate a reasonably high level of accuracy in the use of the targeted form) continue to make errors in the use of the English article system . . .". A case for replication with lower-level students? They arguably have a greater need for correction and may benefit more from the "treatment".
- Furthermore, "L2 writers across English language proficiency levels" had us wondering whether all L2 writers demonstrated this, or whether it might depend also on the participant.
- Does the literature suggest other errors appear across proficiency levels that might be worth testing in any replication?

» Activity 8d: PRACTISE AWARENESS-RAISING: THE TREATMENT AND INSTRUMENTS USED

Now focus on the section entitled "The study: Treatment" and "Instruments" and carry out the same initial awareness-raising reading strategy you saw in 3.2 Routes to Replication: Reading for Awareness-Raising (p.28) above using the space around the text for your annotations. Then use the targeted question sections *Specific Task Variables* and *Instruments/Tools Used* above to help guide your responses.

Once you have completed your own reactions, compare your answers with our suggestions below.

- Groups 1 and 3 both received "direct written CF". Even with a "simple explanation", this kind of feedback will require deeper processing than that offered the other groups. Is this form not better suited to the more advanced student, then? Would the feedback need to be further simplified in any replication with lower-level students . . . or maybe presented in the L1?
- A replication could try out other ways of presenting the written feedback such as marginalized indications of numbers of errors (but no indication of where in the line) and/or rough indications above the offending area of where the error may lie.
- The CF obtained is all written. Might a replication usefully compare oral feedback in at least one group? How might this work?

- Group 4 (control) received no feedback. We wonder how fair a comparison this really makes for any outcomes. Might a replication provide some other CF approach for comparison?
- There appears to have been no measure of error incidence taken. How do we know how many errors each group had to act upon? Is it not conceivable that the control group – with no CF – actually made fewer errors to feed back on?

» Activity 8e: PRACTICE AWARENESS-RAISING: PROCEDURES AND ANALYSIS

Now focus on the section entitled "The study: Procedure" and "Analysis" and carry out the same initial awareness-raising reading strategy you saw in 3.2 Routes to Replication: Reading for Awareness-Raising (p.28) above using the space around the text for your annotations. Use also the targeted question sections *Specific Task Variables, Length of Treatment* and *Instructions and Cues* above to help guide your responses. Once you have completed your own reactions, compare your answers with our suggestions below.

- We wondered whether Group 1's more textually intensive CF would have needed more time to process than the others? Why did they all get the same time for differing types of CF? Might a replication look at this?
- "The delayed post-test was administered 10 weeks after the pre-test. The teachers had agreed to not provide any instruction or correction on the targeted forms during the interim period." But this was an ESL situation: one can assume all participants were actively interacting with the language outside the test situation. Would this potentially affect outcomes? An EFL-in-a-non-English-speaking-environment replication might be a useful comparison replication?
- Do you feel the accuracy rate/calculation should be attended to in any replication? Why?

» Activity 8f: PRACTICE AWARENESS-RAISING: THE RESULTS

Now focus on the section entitled "The study: Results" and carry out the same initial awareness-raising reading strategy you saw in 3.2 Routes to Replication: Reading for Awareness-Raising (p.28) above using the

(continued)

(continued)

space around the text for your annotations. Use also the targeted question section *Statistical Significance and Effect Size* above to help guide your responses. Once you have completed your own reactions, compare your answers with our suggestions below.

- Table 1: At the pre-test we note that the SD (standard deviation) for Groups 1 and 3 indicated a larger dispersion of scores for such a small group. Might a replication try to achieve better homogeneity before treatment? Participants could be increased in all the groups in a replication?
- Table 1: Control group recorded increased SD on delayed test. It would be interesting to see what happened. Replication might provide more detailed distribution data?
- Figure 1: The indirect feedback group (Group 2) seem to be initiating a *de*crease at the delayed post-test stage (admitted later in the text). Might a longer time replication investigate if this persists?
- Figure 1: "However, at the time of the delayed post-test, the participants who received indirect feedback could not sustain this improvement and therefore did not differ significantly from those in the control group" (214). Follow up in a replication? Perhaps a shorter period might reveal how long the improvement could be sustained?

» Activity 8g: PRACTICE AWARENESS-RAISING: DISCUSSION AND CONCLUSION

Finally, focus on the section entitled "The study: Discussion and conclusions" and carry out the same initial awareness-raising reading strategy you saw in 3.2 Routes to Replication: Reading for Awareness-Raising (p.28) above using the space around the text for your annotations. Use also the targeted question section *Does the Generalizability of the Study Need to be Further Tested?* above to help guide your responses. Once you have completed your own reactions, compare your answers with our suggestions below.

- ". . . all three treatment groups (indirect and both types of direct feedback groups) outperformed the control group in the immediate post-test." Again, we wonder how far the comparison is concluded to be a fair one. Maybe the replication could provide the control group with some basic kind of CF as a better comparison?
- ". . . those who received written meta-linguistic explanation and those who received both written meta-linguistic explanation and an oral form focused review were able to retain their accuracy gains across the 10 week

period." Yes, but it seems immediate post-test to delayed post-test results were not as high as pre-test to immediate post-test. A further indication that a longer period of investigation in a replication might prove useful?

- ". . . also be effective in targeting certain types of errors made by advanced L2 writers". We wonder whether L2 writers, by their nature here of being "advanced", are more attuned to, and more receptive to, CF and willing to act upon it? It would be interesting to see if lower proficiency levels react in the same way.
- "For advanced L2 writers, it is clear that one treatment on one error category can help them improve the accuracy of their writing." Is it worth taking this up in a replication? Would a focus on more items improve CF take up or even weaken it because participants now have more to focus on?
- Take up some of the suggestions as replications in the first and last paragraph on p. 216.
- "From a pedagogical point of view, this study reveals that the provision of clear, simple meta-linguistic explanation, namely, explanation of rule(s) with example(s), is the best type of written CF for long-term accuracy." We wonder what actually makes the CF memorable. Why was the direct CF gain not sustained? Replication with only this type in a number of separate treatments and groups over time?

Notes

1 Carter, M., Ferzli, M., & Wiebe, E. (2004). Teaching genre to English first language adults: A study of the laboratory report. *Research in the Teaching of English*, 38.4, 395–419.

2 Plonsky, L., & Derrick, D.J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, 100, 538–553.

3 Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics*, 31.2, 193–214.

4 Norris, J.M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65 (Supp. 1), 97–126.

Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 23–45). New York: Routledge.

5 Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the New Statistics: Estimation, Open Science and Beyond*. New York: Routledge.

6 Oswald, F.L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, 30, 85–110.

Plonsky, L., & Oswald, F.L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 106–128). New York: Routledge.

7 American Psychological Association (2010). *Publication Manual of the American Psychological Association (6th Edition)*. Washington, DC: American Psychological Association.

8 Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G.K. Porte (Ed.), *Replication Research in Applied Linguistics* (pp. 116–132). Cambridge: Cambridge University Press.

4

WHAT KIND OF REPLICATION SHOULD YOU DO?

From the Inside, Looking Out: Initial Critique and Internal Replication

4.1 Deciding on the Approach to Replication

4.1.1 Initial Advice

The literature generally talks of two overarching types of replication: internal and external. The former would see action upon the original set of data presented by the author and is most often carried out by that author, prior to the publication of results and therefore already incorporated into the original study. Unlike external replication, this process will, therefore, not require the study to be repeated in any way. It provides a very useful *re*visit to the study and cross-validation, however, and potentially offers an additional element of reliability in the sense that we see the author providing further analysis and scrutiny of the data – either using the same analysis procedures or others (see below). Such an approach can help us see both whether any errors present at that analysis stage or the outcomes were themselves affected by the nature of the initial analysis undertaken.

As a "by-product", internal replication can also give us – as readers – the chance to re-examine the methods used and conclusions drawn. This kind of analysis should be an essential element, of course, in any peer review of a paper which is submitted for publication. After all, while these days the use of specialized statistical software might give us hope the math operations would be correct, comparatively few referees might request the additional data often omitted through lack of available space, nor re-estimate the models provided in the paper. Decisions about publication tend, therefore, to be made based on more general criteria of the perceived acceptability of the results (see Sokal & Briemont, 1998).[1]

Yet, as we have suggested in the Introduction, replication should be part of the usual way of our going about things. We all make mistakes in planning and

carrying out our research, most of them unforeseen, unwanted, and unintentional. While the recent furor in the press indicates that there can sometimes be a worryingly thin line in published empirical research between genuine error and a desire to deceive, we might see internal replication as a first line of scrutiny (or defense) to both find – and correct – errors. Without such safeguards, we can simply never be confident in the outcomes presented; with any of the many approaches to replication described here and elsewhere taken up as the norm, we might begin to see practices change.

Unlike procedures in the hard sciences, where later scrutiny typically comes from the assumption that external replication will further confirm or disconfirm findings (see Introduction), much of the focus in the social sciences has been on aspects such as the correctness and appropriateness of the statistical methods used, the selection procedures, and/or the parameters or operational definitions used by the author. As a result, however, we will often be perusing a number of studies dealing with similar aspects of an area of AL but using very different assumptions, criteria, and statistical procedures to present their outcomes. Even more reason, then, why each study needs to be judged carefully on its own merits at the same time as we think about its contribution to the wider field of knowledge in that area.

4.1.2 Routine Checking of a Study: Initial Critique

We are first going to assume that neither internal replication nor routine checking has been carried out by the original author; we are to set about the task of reviewing what has gone on. There would, indeed, be little point in attempting to replicate an unsafe study!

Your aim is to become familiar with the study chosen: even as a training exercise, such a procedure can be a useful, initial practice procedure which can help you acquire the kind of insights into a study which will prove useful when you start out on your own external replications of other work. It sees you focus as much on the data presented as on a critique of the way a study was carried out and its conclusions drawn.

Any approach to replication, internal or external, involves data sharing. Any critique of what has gone on means we need to have the full picture! This routine checking stage principally involves verifying the data rather than actually rerunning the experiment but should also mean close examination of what went on to obtain that data. Verifying outcomes has to come from somewhere, of course, and we will need enough detail in the procedures and method to understand what has then gone on to be presented in the results.

The selection of our target study should follow similar lines to those described in the previous chapter as regards dovetailing our interests down from a substantive problem, through a sub-area, to locating and obtaining a possible article, and then confirming whether we have all the data used and necessary to replicate the specific numerical results in the tables and figures presented.

And there we have what may prove an initial stumbling block: can we get our hands on all the data needed to reconstruct the study – in particular, the method and results (including the raw data)? As we have seen in the previous chapter, this may well mean looking in online repositories (such as IRIS, see below), the supplementary online materials published by journals, on research project webpages, or soliciting help from the original authors.

Routine checking sees you first reading through the paper to see whether results can be reproduced using what is available for examination in the article. Once again, we need to adopt our close-reading-and-commenting technique presented in the previous chapter to satisfy ourselves about any problems present in the data available – aspects such as participant selection, possible errors in data coding and/or transcription, or in the kind of statistical approach employed in the light of the assumptions required for such tests.

Ideally, we would like to see an author improve "reproducibility" in his or her work by presenting enough information and data about the study for the reader to understand and evaluate the work, prior to the more detailed discussion of what has been observed in the Results and Conclusions section of the paper. While there are inevitable restrictions on space in journals, we should still expect to see enough of the results to satisfactorily respond to the hypotheses or research questions initially put forward. The appraisal we make here needs to pay special attention to any tables or figures in which outcomes are presented.

Ask yourself whether you see any anomalies or potential confusion in the **graphical presentation of outcomes** which need to be further clarified? Does the information coincide with what you have read in the text? Is there any information which appears to be missing and needs to be obtained to make an adequate appraisal? Are there any unaccounted-for outliers? Do the quantitative values presented seem reasonable?

Take a good look at the columns and rows in the tables of results. Most likely, these will have been labelled with the variables, group designations, and scores that apply to the paper. If we are to analyze outcomes ourselves, we will need to be sure these correspond to terms that have been sufficiently well explained or defined in the text itself. Thus, for example, are any "scores" assigned raw data or "per x words"?; are independent and dependent variable labels satisfactorily operationalized?; do you spot any apparent anomalies in any results and are these taken up in the discussion section? And, while most statistical data is carried out through special software these days, it might be as well to check the math as far as possible!

The **statistical operations** carried out on the data will usually be presented, and we will want to check that enough is available to permit adequate checking to be carried out. **Descriptive data** will often be a first port of call: we would, for example, look to see the basics: how many participants were involved, how were they distributed into groups, and of how many, what scores were obtained by whom, and what measure of central tendency (i.e., mean, median, or mode)

and dispersion (SD, standard error, confidence intervals) had been used (outcomes can be greatly altered by the choice!). The mean is by far that most chosen in studies manipulating interval/continuous scored data; however, if our routine checking is to be prior to any further replication of the paper (see below), we need to remember the measure is highly sensitive to extreme scores (outliers). In a relatively small group, one or two such scores could shift measures such as the mean and the SD, and it would be interesting to find out from the original researcher what was done with such cases. The SD might be equally worthy of attention as part of our routine checking: larger than normal dispersal of scores in a small group will attract our attention – particularly if we are considering the virtues of a subsequent replication involving more participants.

Much has improved in recent years following the calls for more replication studies in our field, and publishing guidelines from a number of professional organizations[2] and journals themselves now clarify the requirements for data availability and online data storage, designed to help supplement what is presented on the space-limited written page. Having said this, at the time of writing, many journals recommend only the uploading of data collection materials to databases such as IRIS (www.iris-database.org). There remains no *obligation* to do so, and while this remains the case you may well still need to depend on the willingness of the target study author(s) to fill in the gaps. That, in turn, will mean first approaching the journal's website to see where such data is provided. If the journal's main concerns do not lie in posterior scrutiny of, and debate on, their papers – and for many this is not a priority – there will be a disincentive for that journal to publish corrections, a disincentive for outsiders such as you to search for and suggest corrections, and – perhaps more insidiously – a disincentive for researchers to be careful when they carry out and write up their work.

Typically, therefore, you will need to prepare yourself for initial lack of precision and/or more of the space-induced lack of detail mentioned above than you might hope for. As you read, you will need to be able to "spot" where something may need further clarification as a result of routine checking. You might also need to be prepared to change your original choice of target study if faced with the impossibility of obtaining the kind of "missing" data, methodology, or procedures needed to carry out your prospective replication.

» Activity 9 ROUTINE CHECKING

Read this extract from a study on which you have decided to carry out a routine check. At certain points (marked ^) we suggest that some important data/information may be lacking. This may need to be sourced online or from the author. Decide what data this might be and write in the columns on either side of the text.

(continued)

(continued)

Research Questions

Does practice in dictation in class improve participants' knowledge of L2 French structures, vocabulary, listening, and reading comprehension?

Participants

Forty-four students of L2 French studying in their first year of French conversation were put into two groups ^, the experimental and the control group. Groups had the same numbers of participants each and were matched on a number of parameters ^. Participants consisted of male and females ^ varying in age with a mean of 18.6 years old ^. At the same time as the experiment was being carried out, participants were undertaking their other university courses, but none in L2 French ^.

Procedure

The conversation class took place three times a week ^ over a period of 12 weeks. The classes consisted of instruction in a number of elements of French pronunciation ^, delivered in the classroom using both researchers ^. Both groups followed this same syllabus for the period and used the same material ^.

The researchers taught the two groups in turns throughout the experimental period ^. The difference between the two groups came from the fact that the experimental group received 60 practice dictations in their classes spread throughout the experimental period ^. These dictations were presented in supplementary material from the books used which were only available to the researchers and increased in difficulty as time went on ^. The group heard the dictations from audio media at the front of the classroom ^ and followed the instructions presented from the teachers' version of the text book ^.

All the dictations were collected at the end of each class and scored by one of the RAs (research assistants) of the researchers who had been trained in this scoring before the experiment took place ^ and missing and misspelt words were indicated in red ink. Texts were then returned to the participants and they were allowed to review the marking and ask questions before proceeding to the rest of that class. At no time were participants told the main objectives of the experiment.

Testing Data

Pre- and post-test data was obtained by using the *Cercle Français* official test which covers all four language learning skills (testing materials, timing requirements, and other criteria can be obtained from the authors) ^.

At the end of the experimental period the participants were all also administered three dictations. The latter were specially designed by the researchers

and adapted from L2 French readers with which the participants were not familiar. The dictations were then graded. ^

Results

Table 4.1 shows the descriptive data for the two groups and skills subsets in the pre-test. While some discrepancies between the two groups can be observed (the experimental group seems better in grammar and vocabulary than the control and the control better in listening and reading), in general it was felt that the two groups were sufficiently similar to proceed with the experiment as statistical data showed none of the score differences (mean or totals) shown as being significant, ^ and therefore it was felt these results supported the matched selection process mentioned above.

TABLE 4.1 Descriptive statistics: pre-test

Pre-test	*Experimental group*		*Control group*	
Sub-test	*Mean*	*SD*	*Mean*	*SD*
Listening	44.55	23.14	47.65	20.37
Grammar	69.75	22.27	64.86	20.58
Vocabulary	49.03	22.94	46.31	27.36
Reading	44.77	28.42	46.11	23.47
TOTAL	52.02	24.19	51.23	22.94

Table 4.2 shows the descriptive data for the two groups and skills subsets in the post-test. There are similar differences between the groups, with the experimental group better in grammar (but not vocabulary) and the control group better in listening and reading, but none of the results were seen to be statistically significant ^.

TABLE 4.2 Descriptive statistics: post-test

Post-test	*Experimental group*		*Control group*	
	Mean	*SD*	*Mean*	*SD*
Listening	55.66	19.15	57.96	19.43
Grammar	71.53	20.95	70.04	22.86
Vocabulary	64.13	22.88	67.69	23.98
Reading	56.84	23.09	60.78	24.44
Dictation 1	55.16	18.07	56.78	18.08
Dictation 2	20.98 ^	7.75	20.43 ^	7.44
Dictation 3	60.34	19.31	57.72	22.31
Total dictation	45.49	15.04	44.97	15.94
TOTAL	62.04	21.51	64.11	22.67

(continued)

(continued)

It is clear, therefore, that the extra dictation practice with the experimental group did not improve test performance in any significant way.

Conclusions

The use and practice of dictation was not shown to improve the language skills of the experimental group participants. A number of possible explanations can be put forward for this. First, it may be that the instruction ^ had no effect on the participants' abilities or they did not have enough time to exercise their skills. Second, it may be that the failure to show improvement could be due to the fact that participants did not understand a number of the words in the dictations. Post-instruction interviews ^ with randomly selected participants from the experimental group revealed that little extra French language reading was being done outside the classroom and this, too, was not seen as conducive to improvement.

Descriptive statistics tell us only part of the picture: group characteristics will vary across many research contexts and conclusions drawn from descriptive data may only be informative, but not conclusive. We would need to carry out many replications in these contexts and with different participants to come close to checking the validity of the outcomes. **Inferential statistical procedures** are more often called upon as they provide more insight into what has happened and potentially allow us to generalize to other contexts.

There is not space enough here to discuss all the current inferential statistical procedures you may come across in your reading of papers, and whose assumptions you will need to check. Rather, we present below a general discussion of the kind of things your routine checking phase should cover, with particular regard to the immediate job at hand – checking what has been reported as far as we are able to – and establishing where replication of the study might help clarify outcomes or move things forward.[3]

We would initially want to make sure the various statistical operations carried out on the data are made clear and, particularly, the assumptions such as normality of distribution or selection procedures. Any subsequent replications of the paper will need to take apparent omissions into account (how far, for example, can we accept the wish to generalize from these results if these assumptions have not been met?).

As we mentioned in the previous chapter, our initial attention can be drawn to the **statistical significance level**. There are a number of specific discussions already mentioned which are worth reading about the perceived importance of significance level testing,[4] but for our present "replication objectives", a number of general points should be borne in mind.

First, as we digest the outcomes, remember that the significance level chosen by the researcher is unlikely to have been selected through any reasoned argument based on previous, similar research. Depending on the cut-off point previously adopted (α level), we could find ourselves presented with results which rest too heavily on the rejection of the null hypothesis. And then, depending on the replication interest generated by the study, there might be a case to be argued at the very least for strengthening the original significance level assigned to the data outcomes.

Next, you will need to bear in mind that the significance statistic you read will relate closely to the **sample size**. While large numbers of participants might be a positive sign as regards the external validity of the sample, one can obtain a statistically significant outcome if your sample is large enough with even a small effect. We have also noted above the tendency for journals to publish papers wherein null hypotheses are rejected, but failure to do so is arguably even more interesting from the point of view of the emergent replicator. Along these lines, the number of statistical tests should also be considered in light of statistically significant findings. It is not uncommon, for example, to find literally dozens of statistical tests run for a single study. This is not a prudent approach as doing so greatly increases the chance of Type I errors.

We have warned in the previous chapter, however, about reading **too much into, or beyond, the significance level** presented. The outcome does not present us with the chances of the results being replicated – however "statistically significant" they may be. As Nassaji (2012) points out, part of the misunderstanding can come from a misinterpretation of what significance testing can show us – and what it cannot. Significance testing "only" tells us about the probability of the outcome, assuming the null hypothesis is correct. It does not tell us "the probability that a certain mean difference . . . can be *repeated*" (Nassaji, 2012, p. 101, our emphasis). Nassaji also cites work by Cumming (2008)[5] in which the latter – through a simulated set of repetitions of experiments – found that the *p* value varied so wildly that it is clear this value provides no proven indicator of generalizability, let alone replicability.

As a way of strengthening your own internal replication and, by extension, the external replication that will be carried out, the bulk of your attention should be on the effect sizes which express the strength or magnitude of an effect or relationship (see below). The effect sizes indices most common in L2 research are Cohen's *d*, *r* (the correlation coefficient), r^2, R^2, eta^2, and the odds ratio (for categorical data). Although the reporting of effect sizes has increased dramatically in the last 15 or so years (Plonsky, 2014),[6] it is still commonplace for them to be omitted. In many such cases, you can calculate the effect size yourself. The formulas are quite straightforward and involve only the most basic arithmetic based on descriptive statistics such as means, SDs, and sums of squares (in the case of ANOVA). An understanding of the effect sizes will assist in arriving at a clear and comprehensive understanding of the initial study in question, will help make sense of your own results, and will also provide a basis for determining an appropriate

sample size for the replication (via a priori **power analysis**, see Larson-Hall, 2016 for more information). In Null Hypothesis Significance Testing (commonly NHST) (see Cumming & Calin-Jageman, 2017),[7] sufficient statistical power is required to help ensure "real differences are found and discoveries are not lost" (Larson-Hall, 2016, p. 156). Statistical power can be calculated relatively easily with online calculators (e.g., www.danielsoper.com/statcalc/) which usually require you to provide information about (a) the statistical test you will use (e.g., multiple regression, ANOVA), (b) the alpha level or statistical significance level you will use (e.g., 0.05), (c) the desired effect size, and/or (d) the desired statistical power level (usually greater than or equal to 0.80). See Larson-Hall (2016) for more information.

Finally, we return to a routine check on the assumptions for the statistical procedures used, and will use some of the more popular procedures to be found in AL to illustrate the points.

4.1.2.1 Checking for Test Assumptions: Chi-Square

Chi-square addresses the relationship between two nominally (or frequency) measured data. A popular test in many AL papers, chi-square will assume key assumptions have been met previously. Independence of groups and observations means that each participant should only be seen to contribute data to only one cell of the chi-square table. Similarly, however, we would want to make sure that one participant is not seen to present considerably more data than others in the same cell. You would need to look at the research question and data collection carefully to see how far this might be likely: for example, in a hypothetical test of relationship between L2 students amount of learning strategies measured and success in assessments in language learning, our experience might suggest successful students might well normally have acquired and be using more learning strategies than less successful ones. You might also come across frequency-measured data being obtained from the same group on a pre- and then a post-test measure. The data then comes from the same participants, albeit at different points in time, and the independence of observations assumption is not met.

As with other statistical tests below, if we felt the study still merited replication, there are often alternative routes to help us if the relevant assumption is seen not to have been met. In the above-mentioned situations of lack of independence, for example, Larson-Hall (2016)[8] recommends alternatives such as a binomial test when there is data potentially in more than one cell, or the McNemar test when using repeated measures and categorical data outcomes.

4.1.2.2 Checking for Test Assumptions: Correlation

Our routine check will need to ensure correlation is not used to conclude anything beyond that procedure's main purpose. In correlating two continuous

variables, for example, the researcher will be looking for some kind of positive or negative relationship: it might be that there is a (positive) relationship observed between being able to speak and write a foreign language such that the better you are at speaking, the better you are also at writing that language. It would not be correct to go on to conclude that the two variables have some **causal link**, that better L2 speaking somehow brings about better L2 writing – although subsequent research may indeed try to confirm such a possibility.

Again, as part of this routine check, we would want to check that the assumptions applied to this procedure were accounted for satisfactorily. We would expect there to have been examination for **normal distribution** and **homoscedasticity** (i.e., having equal statistical variances) of data: more often than not this is assumed rather than established although most statistical programs will estimate it satisfactorily. Ideally, there will have been some examination of the data visually to establish that the data is **related linearly** – the data relationship once plotted should approximate to a straight line. Furthermore, the data or scores from one participant should not impact those of another – some apparent effort to have **randomized data-gathering and group assignment** (i.e., to control or experimental) will be reported.

Replications might equally derive from our checking of the SD presented, since the **bunching of high or low scores** might indicate how far a parametric correlation procedure was warranted. Similarly, we would want to review the logic behind the correlation tested: if, for example, we were presented with an experiment correlating L2 listening comprehension with age, we would want to make sure that the participants' age range was indeed enough to permit a wide enough set of data to begin with. (A restricted range in one or both variables being correlated can attenuate or reduce the observed correlation between the two variables.) If not, a replication might want to widen this further.

4.1.2.3 Checking for Test Assumptions: T-Tests and ANOVA

Arguably, the *t*-test, in its many representations, is one of the most commonly occurring procedures you will read about in AL experiments. Just as commonly, however, its use has led to abuse, and it would be a sound move for our routine checking of papers both to attend to any examples we find and note how replications might be used to clarify any doubts or provide more useful evidence for or against any differences manifested. The independent-groups procedure, for example, compares the means of two groups and presents the statistical significance level at which the researcher can determine a difference between the groups due to some variable other than chance.

First, the data used should be seen to be **normally distributed**. This would usually have been checked in any statistical program used – assuming there was enough statistical power residing in the test itself – but some initial insights can be gained by reviewing the apparent distribution of the data and the SD presented.

We would expect the researcher to **identify the specific test used**, not least because of the different assumptions which may apply to each. Non-parametric tests such as the Wilcoxon or Mann–Whitney U are available to the researcher who finds the normal distribution of data is not confirmed. The second assumption, of **equal or similar variances** (SD^2) or the homogeneity of variance, is one which is often assumed rather than checked in much of what you might read: in our routine checking, it might just be a matter of logic, too. If, for example, a paper is comparing the error marking of two groups of L2 English teachers, native as against non-native, we might feel the former were less likely to present the kind of variance which the latter group might demonstrate. Once again, however, a replication – if it is to be carried out or recommended – can use additional estimating procedures to account for such inequality. Third, we would want to see the data measured as **interval or score** measurements. Any conversion of data from nominal non-continuous to continuous score data (such as frequencies to percentages scores or raw frequencies to rates) in an effort to accommodate the *t*-test would need to be satisfactorily explained or justified.

The same goes for conversions of continuously measured variables to categorical ones. This transformation, which you might see employed to enable a comparison of groups, is ill-advised for the loss of variance it entails. Imagine a study in which the researcher was interested in examining the relationship between motivation and proficiency and, to that end, gave a group of L2 learners both a motivation measure and a proficiency test. Some researchers might be tempted to divide the motivation scores at the median to form two groups whose proficiency scores would then be compared on a *t*-test. This approach is appealing in that it would provide an estimate of the difference in proficiency between low and high motivation learners. However, we strongly advise against this practice on several accounts. First, as mentioned above, this approach results in a loss of variance. An intervally measured variable has been converted into a dichotomous one and, thus, a lot of information about participants' motivation has been discarded. Second, unless there are theoretical grounds to believe that motivation exists as a "low" or "high" phenomenon, the grouping is not sound. And third, the more appropriate – and, actually, simpler – approach here would be to run a correlation between these two variables.

The main assumptions for one-way and factorial ANOVA – which tests a comparison of three or more group means rather than only two groups – are the same as those for the *t*-test. However, since we are now reviewing the data – and specifically the mean scores and the dispersion around those means (i.e., SDs) – from more than two groups (unlike the *t*-test) and levels (in factorial ANOVA), the researcher will need to go further into the outcomes to establish exactly where the differences between the groups lie.

4.1.2.4 Checking for Effect Sizes: Principal Considerations

A number of authors, in the field of AL and others, have pointed to the limited information provided by NHST and their associated *p* outcomes,[9] together with a large number of issues which highlight their limitations and weaknesses.

The practical significance of findings – presented through an appropriate effect size statistic – is nowadays required by more and more AL journals. One noble objective behind this is the desire to see experimental research as a pooling of findings rather than mere accumulation of disparate outcomes. It is reasonable, therefore, to expect to see the researcher discuss the perceived importance of their findings and the effect(s) observed or calculated compared to previous work.

For the potential replicator and – now – routine checker, these effect size statistics are one of the most important pieces of information to focus on in the "Results" section of papers beyond any NHST outcomes or *p* values. Knowledge provided about the effect size can determine the nature of the statistical power to be assigned in any subsequent replication. It can also help us at this stage to check what might have accounted for the outcomes reported. As we shall see below, however, the journal's requirement to include an effect size statistic can also lead to results being presented in a generic form with little or no discussion of the research – past or present – and the practical implications of the outcome.

The effect size is a simple and readily interpreted measure, but it can also be sensitive to a number of spurious factors, so care is warranted as *we* interpret it – particularly across similar studies. Plonsky and Oswald (2014)[10] lament that the reporting of effect sizes has become a mechanical product of "the check off list in author submission guidelines" and post-results interpretation is at a premium. Nevertheless, such explanation is an important signpost for any of us interested in replication. In their discussion of the importance of meta-analyses for the field, Plonsky and Oswald provide some useful guidance for our interpreting effect sizes which can be used for your potential replication (e.g., interpreting effect sizes based on meta-analytic work in the field of study instead of using broad-brush benchmarks):

A. What do previous, comparable studies tell us about the outcome?

"Comparable", in a scenario where few replications are being carried out, is more likely to signal studies which have measured the magnitude of similar relationships, or meta-analyses which have measured this across a number of similar studies in the main, or sub-domains of the study in question. Here, two caveats. When we are looking out for cumulative evidence for things like effect sizes, and in any comparison of studies, there are two important, if obvious, cautions. These are particularly relevant if we are going on to plan possible replications of the studies in question.

First, as with any comparison, we need to know that we are **comparing like with like**. If not, can we combine them, average them, or even begin to compare them? Thus, if we are looking at two studies, both of which test whether a particular approach to L2 vocabulary learning is more successful with intermediate than advanced learners, but wherein different effect sizes were reported, we would need to see whether constructs such as "L2 vocabulary", "Intermediate (level)", and "Advanced (level)" were operationalized similarly. Then, we would need to see how the treatments were carried out in terms of **procedures**: for example, if the treatments were the same or similar but the time intervals between the pre- and posttests were significantly different, any calculated average between the effect sizes would have little meaning.

B. Interpret the mathematical facts.

The effect sizes you read will have a numerical outcome which can be readily interpreted based on meta-analysis and field-specific benchmarks. As we will see below, however, you also need to be careful when undertaking such a rigidly-based interpretation in such a person-variable field as AL. In many cases, meta-analysis in a particular field of study (e.g., pronunciation instruction) can be the most helpful.

Thus, Cohen's *d*, for example (see below for more details and the *r* family), will often be interpreted (and accepted) as "*0.2 and below = small effect size; 0.5 = medium effect size; 0.8 and above = large effect size.*". In fact, Cohen himself warned against **too inflexible an interpretation** across studies of these statistics, and we would therefore need to be wary when looking to replicate the study in question using the effect size reported as a possible reason. An effect size of 0.60 might be interpreted as as "small" in the context of one study, while in another context 0.60 may well be considered to be "large". Similarly, the same adjustment of interpretation might be needed if a small effect can be seen to be meaningful.

Therefore, treat these assimilated descriptors with caution as we routinely check papers: if necessary for a potential replication which we want to undertake, consider consulting the author about his or her interpretation of an effect size beyond the number reported. And remember, we would be wrong to assume a "large" effect reported is, *ipso facto*, more important than a "small" one. We would need to consider also the practical significance of the outcome as well (see below).

As we mentioned above, comparisons need to be used with care as the basis for future replications. The diverse features of research studies we have read about in our first chapters can make a smaller effect size from one measure appear more important than a greater effect size based on another type of measure (see below). Therefore, the moral of the study is not to be over-influenced by the reported magnitude of the effect size statistic: much depends also on your own critique (and that of others) as regards the strength of the research methodology itself!

A final consideration here would be the **typically underpowered studies** often found in L2 studies, where we might read of small participant numbers and/or less-than-ideal selection procedures. Effect sizes in such scenarios might well be overstated in the circumstances and in need of further confirmation.

C. Consider the possibilities of the effect size statistic remaining at this level over a longer period or amount of time.

One of the questions a potential replication might take up is to investigate how far an effect remains as strong (or as weak) **over longer periods of time** than were studied in the original paper. Further information about effects over time – whatever the outcome – is only likely to contribute to the body of essential knowledge about a key study as we gain more insight into the practical significance of the findings. A replication can test these in various ways in a replication: we might use the same treatment over a longer period, we might use a reasoned variation of the treatment over that same period, or even experiment with another way of measuring that treatment over time. Similarly, we might want to study how far the intensity of time in terms of hours per session might impact outcomes.

D. Consider the possibilities of the effect size statistic changing along with the contexts in which the research takes place.

As Plonsky and Oswald suggest, **research setting** may also explain variability in outcomes, and may help us envisage another way in which a replication may help clarify the result from the target study. These authors specifically single out two possible context changes which might affect outcomes (the typically tighter controlled "laboratory" vs classroom; ESL vs EFL), but you will doubtless think of others during your reading of the target study (see below, Approximate Replications). A further context variable mentioned here by the authors concerns **the true nature of the control group** (treatment/no treatment) when this is used to calculate the effect size. In most cases you will be told the control group received no treatment, but elsewhere local stipulations or timetable needs might dictate that this group received a "traditional or alternative" intervention. Again, the potential replicator may want to tease out the true nature of any experimental effect by comparing treatment/no-treatment control group outcomes.

These authors also remind us that effect size outcomes need to be weighed against **specific methodological aspects** of the study in question. Aspects such as reporting of reliability and validity measures for the dependent variables ". . . pretesting and delayed posttesting, random assignment to experimental groups" might conceivably affect the strength or weakness of the statistic reported and, as such, be a useful focus for replication.

4.1.2.5 Checking for Effect Sizes: Chi-Square, Correlation, and ANOVA

Given the fact that many AL journals are only relatively recently requiring the reporting of effect sizes in submissions, you will need to be wary in the comparison of effect sizes themselves. It would not be surprising to find earlier work without a reported effect size or – at best – an unspecified one but no detail about how it was calculated. All is not lost, however. If the basic statistics of number of participants, mean, and SD are available, a good estimate of the effect size can be made and a number of online calculators can help.[11] Such calculations will also help the potential replicator decide on the ideal sample size to extend the power of a study in any further work.

As we discussed previously (p. 55), effect sizes go by different names although they essentially do the same thing. You will most likely read of members from two main "families": *d* (standardized mean differences) and *r* (strength of association) and – depending on sample size – associated corrections for possible bias.

In chi-square, the statistic will indeed tell the researcher how well two variables "fit together", or how well they are related, but *our* interest will be more in the strength of the relationship presented. A result can be statistically significant but whether that significant difference is of practical concern is another matter. Effect sizes either measure the sizes of associations or the sizes of differences. The statistic you read will be reported according to the effect size procedure chosen.

In one of most common correlation statistics, the Pearson *r*, this essential effect size will be presented as R^2 (r^2) although *r* itself can also be considered an effect and can be interpreted on its own as well. Much may be read into this effect size so it may be worth stopping and reviewing it – particularly in the light of other similar studies in the area in question: *in general*, for the field of AL, a statistic of 25% or more here can be thought to show a strong effect and below 6% a small effect. However, common sense might also need to come into the interpretation – as with any effect size statistic (see below). What remains outside this statistic (i.e., the remaining 75% in the case above) is presumably due to some other variables than those studied, including measurement error. A replication might want to delve further into this unaccounted variance.

Both the *t*-test and ANOVA should see an effect size reported, and once again the researcher who is interested in adding to the useful information provided by the study through a replication might find here an interesting starting point for discussion. A statistical significance statistic (*p*) will only tell us so much about the difference(s) between groups: with an eye to the need for replication, it does not tell us what we really also need to know: the size of the effect. Reporting the effect size, perhaps together with an estimate of its likely "margin of error" or "confidence interval", will be essential.

In ANOVA, we might see two effect size statistics reported (there are more!), as an omnibus and breakdown of comparisons. The latter – calculated in the same

way as for the *t*-test – will be of greater interest to our cause in that the outcome will give us a sense of the size of the effect of interest. Eta squared (η^2) is used in ANOVA, where each categorical effect has its own eta squared statistic, so we are given a specific measure of the effect of the variable in question.

As we mentioned above, the effect size is particularly useful as we check the data in a paper and consider how a replication might shed further light on the outcomes: it quantifies the effectiveness of a particular intervention, relative to some comparison. It moves us along from a basic question such as "Does the intervention work or not?" to the far more useful, "How well does it work in a range of contexts?". In a language learning scenario, for example, we might undertake a replication to try to increase the effect reported in the original study, and thereby try to show that even a relatively inexpensive change might better academic success through an increase of only 0.1 in the effect size. In that context, and particularly if a lot of students were positively affected, this could be a significant outcome through a replication of the original. Moreover, by placing the emphasis on the most important aspect of an intervention – the size of the effect – rather than its statistical significance (which conflates effect size and sample size), it promotes a more scientific approach to the accumulation of knowledge. Thus, as you pass through this routine checking approach, it is as well to keep in mind that a statistically significant outcome is often inconsequential. Statistical significance is affected by the sample size as well as the effect size. Thus, when checking a study with a view to possible replication, it is as well to remember that – moving onward – practical significance may well need to be shored up by working on the effect observed.

In practice, you will more likely see partial eta squared (ηp^2) quoted in papers. This may be down to the fact that the classical eta squared has a disadvantage that as you add more variables to the calculation the proportion expressed by any one variable tends to decrease, making it more difficult to compare the effect of a single variable across different studies or replications. For smaller samples and one-way ANOVAs you may see omega squared (ω^2) being used. In general, omega is considered a more accurate measure of the effect, where $\omega^2 = .01$ is considered a small effect and $\omega^2 = .06$ and .14 are considered medium and large effects respectively.

One further statistic you might want to look out for – together with the effect statistic – and one which will help us interpret the outcome even more *and* feed into any suggestion for replications is the *confidence interval*. As Plonsky and Oswald suggested above, an effect size statistic from a large sample is probably going to be more accurate than that from a small sample. The addition of a confidence interval statistic adjusts for this margin of error and offers the same information as in a test of significance: thus, a "95% confidence interval" is equivalent to taking a "5% significance level" but using the former keeps the focus on the effect size rather than the more problematic *p* value. Pooling the information from an effect size and the associated confidence interval will additionally help you to evaluate

the relationships within data more effectively than the use of *p* values, regardless of statistical significance (see Cumming & Calin-Jageman, 2017).

If replications of a key study are undertaken – on an individual basis or as part of a research program by a group of researchers – and a resulting set of effect sizes obtained, the different sizes can then be collected to produce a better estimate of the average size of the effect being considered. While replicating a study can obviously help in providing more "evidence" for the strength of an effect, the subsequent averaging of effects from a number of replications will inevitably include studies with "large" as well as "small" effects. The "average" effect size then presented will not tell the whole story of what might have accounted for the differences in each study's effect size in the first place, and our undertaking of a routine check of the individual studies can provide a useful contribution. Our routine check might enable us to come up with an interesting observation in the case of a study with few participants such as an intact class, for example, where there might be a small effect registered in a statistically *in*significant result. Such an outcome may not be of immediate interest because of the original statistical (in)significance outcome, but the results of a number of replications of such experiments combined *are* likely to be statistically significant. In terms of replicational strength, this cumulative evidence will also be of greater scientific *and* practical weight as it will have most likely been consequent upon a number of different contexts and perhaps from a number of everyday settings.

To this same end of potential replication, it is also worth noting whether an effect size is reported in the case of the reporting of *non*-significant outcomes, too. Even in non-significant contexts effect sizes with confidence intervals can indicate to what extent the non-significant findings could be due to inadequate sample size.

» Activity 10: LOOKING AT EFFECT SIZES

Look at the "Results" section from the paper by Carter *et al.* (see Introduction, p. 10).

a. Check above – and in one of the statistics guides mentioned above (p. 57) – to see if the assumptions held for the ANOVA tests carried out.
b. The paper uses an ANOVA procedure but does not present an effect size statistic. Decide if – from the information available – you can calculate any of the effect size statistics mentioned above from the online calculators available, for example, at www.psychometrica.de/effect_size.html#anova. If not, describe the kind of information that would have been useful for the authors to provide.

» Activity 11

Choose *four* experimental studies in your area of interest where use is made in the analysis of a *t*-test, correlation, or ANOVA. Once you have read them thoroughly and noted down the research questions/hypotheses, focus in on the *results section of each – particularly any tables – together with any relevant discussion* that follow and note down:

a. the groups being compared;
b. the statistical procedure(s) carried out on the data;
c. the alpha and statistical significance level(s) reported;
d. the choice of effect size statistic and the outcome.

Then, *first*, think about why the statistical procedure might have been considered the most appropriate and what alternatives there might have been.

Second, decide whether the alpha and significance levels were acceptable and whether they took their lead from any previous work.

Then, consider whether the effect size statistic was the best choice possible and decide whether it has then been satisfactorily considered in the discussion of results.

Finally, and based on your considerations, think about how you might use the evidence presented *through the effect size statistic* to justify replicating the study.

4.2 Internal Replication: Cross-Validation, Jackknife, and Bootstrapping

Before we move on, in the next chapter, to ways of replicating a study yourself using external replication, it would be wise to indicate the kind of internal checking that the researcher him or herself can undertake. This kind of replication, therefore, does not involve redoing the study.

As this internal process is one the researcher will want to perform on his or her own data before publication, we will briefly describe the most common options here so that you are adequately able to assess and critique any such operations as you might read them.

It should be emphasized at this point in our reading of papers – as we are still actively on the lookout for a suitable study to replicate – that the execution of internal replication is not an alternative to external replication: the former does not obviate the need for the other. It is simply a way for the researcher to present his or her results and conclusions with more confidence in their *potential*

replicability and without setting up another sample and actually redoing the study. As we see below, all three approaches to internal replication currently in vogue have their disadvantages, too.

All three typical approaches to internal replication would see the researcher using the original data and through subsets in separate samples which are then reunited in distinct ways. In itself, this can also help the researcher, therefore, to do the kind of self-critique of his or her method and conclusions we have been looking at above. All the procedures below involve the researcher in repeated sampling of subsets of data from the overall pool to present replicate samples. The statistic needed is then calculated for each replicate and the SD is calculated across all the samples to arrive at the standard error of the estimate.

Cross-Validation

This process looks to resample the data and investigates the predictive power of the specific statistical procedure, by comparing the outcomes of one subset against those of another. To do this, the total sample is divided randomly into at least two parts or splits (a training sample and a cross-validation test sample) and identical analyses are carried out on each one. Then, one after another, each part is analyzed and fitted to the test set. This gives you a large number of estimates of prediction accuracy which can be combined into an overall measure. The objective is to discover whether the result is replicable or just a question of random variabilities.

As you might expect, however, the outcome can vary a fair amount depending on what data or observations are in the training split and what are in the validation split. Ideally, we would want to be sure that both training and test samples are sufficiently large and diverse to be representative of the whole data. However, if the sample size is small, logically each analysis is performed with a smaller number of observations and care needs to be taken in interpretation.

Jackknife

In this procedure we once again resample different subsets of the original data. However, this then becomes a step beyond cross-validation as the same test is repeated by omitting samples, one at a time. Recalculations are then carried out on the remaining data. In this way, an attempt is made to partition out the effect on the remaining sample of a particular case from the data, and we can increase the confidence in the potential replicability of outcomes by testing the variability of the remaining samples. The jackknife estimate is the average of values obtained across all the samples (see above). This is a particularly useful replication procedure when the data dispersion in the distribution is seen to be somewhat wide and/or extreme scores are evident.

Once again, however, the procedure has been called into doubt with small samples: sample size itself imposes a limit on the number of subsets that can be

obtained – a problem across many of the resampling methods involved in internal replication. In the jackknife, too large a group can itself lead to larger numbers of combinations being needed. Furthermore, as analysis continues over the same sample, any original sampling errors effects might be magnified (Nassaji, 2012).

Bootstrapping

In comparison with the above approaches, this method is often thought to be a more thorough approach to internal resampling. It consists of copying the data sample a large number of times into a large mega-file of data. Thus, many samples are drawn from the "mega" sample, and the results are calculated for each one and finally averaged, standard errors are calculated, and confidence intervals also averaged. The theory behind this is that we make use of the whole sample to represent the population and many samples of the same size from the original sample.

One of the advantages of this technique over the above is that this procedure does not delete individual data or create unpredictable splits but rather makes use of all the data at once and presents different combinations of the sample. Furthermore, in jackknife and cross-validation the number of observations/data in the sub-sample selected is smaller than the original sample while in bootstrapping each reanalysis involves the same number of observation/data as the original sample.

Our conclusion must emphasize not only the usefulness of internal replication calculations being made, adding as they do, an important element of confidence-building in one's data, but also their limitations as being in any way the "last word" on a study's replicability. While internal replication takes an important step beyond "simple" NHST, external replication – to which we turn now – remains the more reliable way of finding out how far new data, new participants, or new research contexts reflect on the stability of the original study's results.

Notes

1 Sokal, A, & Briemont, J. (1998). *Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science*. New York: Picador. (cf., *The Economist*, October 19, 2013: 28).

2 The tenure and promotion guidelines issued by the AAAL (American Association for Applied Linguistics) now refer explicitly to replication, suggesting that evaluation committees consider "... that high quality replication studies, which are critical in many domains of scientific inquiry within AL, be valued on par with non-replication-oriented studies" https://www.aaal.org/.

3 Readers will want to consult the many statistics books now available to those working in the field of second language acquisition and AL for more detailed advice on these procedures and, in particular, consider what the use of such procedures has assumed in the methodology used and, therefore, the implications for potential replication studies (see, for example, Porte. G.K. (2010). *Appraising Research in Second Language Learning.* Amsterdam: John Benjamins Publishing; Larson-Hall, J. (2015). *A Guide to Doing Statistics in Second Language Research Using SPSS (2nd edn)*. New York: Routledge.

4 E.g., Nassaji, H. (2012). Statistical significance tests and result generalizability. In G.K. Porte (Ed.), *Replication Research in Applied Linguistics* (pp. 92–132). Cambridge: Cambridge University Press; Kline, R. (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington DC: APA.

Norris, J.M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65 (Supp. 1), 97–126.

Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 23–45). New York: Routledge.

5 Cumming, G. (2008). Replication and *p* intervals: *P* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3, 286.

6 Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, 98, 450–470.

7 Cumming G, & Calin-Jageman R.J. (2017). *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. New York: Routledge.

8 Larson-Hall, J. (2016). *A Guide to Doing Statistics in Second Language Research Using SPSS and R*. New York: Routledge.

Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *Modern Language Journal*, 101, 244–270.

9 See, for example, Nassaji, H. (2012). Statistical significance tests and result generalizability. In G.K. Porte (Ed.), *Replication Research in Applied Linguistics* (pp. 92–132). Cambridge: Cambridge University Press.

10 Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64.4, 878–912.

11 For example, www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html.

5

WHAT KIND OF REPLICATION SHOULD YOU DO?

From the Outside, Looking In

5.1 External Replication

5.1.1 Defining External Replication

As we saw in the Introduction, the common tendency in an AL empirical research cycle is for studies to be identified, proposed, designed, carried out, results analysed, and then published, in similar areas or subject matter but then differing considerably in their procedures, instruments of measurement, contexts, nature of participants, and so on (see section 1.1). Consequently, the accumulated knowledge coming out of these studies is equally diverse.

It is a scattergun approach to research: with a sufficient number of studies we might hope to find somewhere amongst the dispersed and fragmented results enough evidence to construct further theory and hypotheses for future research. However, there is no real sense of continuity or systematicity in such a procedure.

Another concomitant problem identified in the Introduction is that this crucial juncture between the dissemination of research and its producing the identification of the next area of study is often seen as one which of necessity extends or follows up the previous research into novel areas or new learning contexts.

Such a procedure typically makes use of the earlier methodology to produce *new* data from *new* contexts with *othe*r participants. The researcher might thereby decide to augment our knowledge with data from a new experiment which has been stimulated by what went on before. The principal aim is *not* to consider the previous study's procedures and outcomes or generalizability in the light of the new data. In such an *extension* or *follow-up* study, comparison between the two studies would be peripheral to the objective.

However, in replication, the crucial juncture is precisely to be found between "publishing the results" and "identifying the (next) research area". In a replication

approach you will add value *by focusing precisely on that previous* step in the cycle. The stimulus for the next step (and the focus of its findings) comes from *one* earlier study, rather than a whole set of studies. The aim is an eminently comparative one focusing on what the replication study says about the original.

This crucial difference between an extension/follow-up study and a replication is an important one – but is not always clear. There should be an obvious and immediate difference of focus and interest evident not only in the very title of the research paper (where we would reasonably expect a term such as "a replication of . . ." to appear), but also in the way the abstract or introductory paragraphs describe what has gone on. You will need to be careful in your reading, as some studies labelled as "replications" are actually follow-up studies which take their lead from previous work but, crucially, do not aim to provide the essential comparisons between the two studies.

» Activity 12

Read carefully each of the following abstracts from papers and decide for yourself, and justify, whether a replication or extension/follow-up study appears to have been envisaged.

1. The research I conducted consists of a 10-year pre/post survey analysis of undergraduate opinion regarding international teaching assistant classroom instruction at a large, post-secondary institution in Canada. It is a replication with minor adaptations of a study conducted at another Canadian university some years before (Author, 1992).
2. This paper investigates how groups of self-selected and teacher-assigned pairs present their collaborative writing. The original research was conducted in New Delhi on female intermediate learners whose L1 was Persian while we are going to replicate it in China on male learners whose first languages are Chinese, Japanese, and Persian. Unlike the original study, our participants will be upper-intermediate students within the age range of 16 to 20 who started learning English together at the age of 6 or 7 in the same school. We used the same methods as the original study but use a locally-written text book to show how successful collaborative writing can be in our teaching context. Our conclusion is that collaborative writing, to be most successful, needs to be introduced into Chinese L2 classrooms from a much earlier age.
3. This study was an approximate replication of Author (1987), which explored how a specially-designed language-learning application improved L2 German freshmen output over a period of one year. Results showed that regular (daily) use of the app resulted in statistically significant improvement in L2 written output but had negligible effect on their spoken German. This research compared both the use of a technological

tool for language learning with traditional teaching. This research tool used was a specially designed app which adapted the original app for L1 French students learning L2 English. The original study failed to consider both technology *and* non-technological language learning tools when gauging student usage and perceived benefits. The results therefore inherently favored technology-enhanced tools and support, when non-technological equivalents may have been just as highly used or effective.

4. Author (1998) studied the way students modified their L2 English interaction in class by means of a number of qualitatively different language learning tasks. This study uses the same methods of data collection and data analysis of the original study, while testing the outcomes using a different L2 – in this case French as well as an enhanced version of the analysis procedure. The results of the replication study partially confirm Author (1993) results, but also indicate that a number of other factors may have affected the outcomes. As a result, the original study's conclusions regarding the extent to which laboratory findings can be transferred to the classroom need to be further investigated.
5. Author (2010) described an extensive reading-into-writing methodology that significantly improved the L2 composition skills of a culturally homogenous L1 Chinese group of EFL college students in China. This present research takes up the same treatment to see if it is as effective with more diverse groups of ESL university students in the United States. An initial pilot study suggested the instrument needed to be further enhanced, and this version was then used over a longer period of time (one year against 6 months in the original). The results indicated gains for the treatment group. Effect size measures underlined the significance of these outcomes. Our conclusion is that a further development of the original methodology is very effective with a more diverse group of learners.

5.1.2 Degrees of Replication: Close, Approximate, and Conceptual Replication

The approach we have advocated in this book has been to regard replication as a natural stage in the research endeavor, and one which is reached after acquainting oneself fully and critically with the target study's aims, procedures, outcomes, and conclusions. Ideally, by having got "up close" to the target study, you will have become sufficiently familiar with it to be able to reflect on its execution and perhaps identify some issues the original author may not have wanted or needed to consider, yet which might yield a different outcome to that observed. Nevertheless, and despite the critical address we have engaged in previously, the replication itself does not have as its aim to reveal potential errors made by the original researcher – although these may become apparent as a result of our work.

Now we have arrived at this stage of familiarity with the study in question, you are likely to be in an ideal position not only to verify what has gone on through your routine checking and critique, but also to conjecture how the results might have come about. It is therefore only a short step from this point to the consideration of how you might usefully build on these results in a way in which you can make a contribution to the field by proceeding to some kind of replication of the study.

As we have seen, there is often confusion about what a replication study does, and what an extension or follow-up study does. Historically, moreover, the literature has referred to degrees of replication according to what is replicated and how, and then defines – in rather too many ways sometimes – how each degree of replication might be termed (see Polio, 2012).[1] This confusion is not new: in a review of papers in the *Social Sciences Citation Index*, Bahr, Caplow, and Chadwick (1983)[2] concluded that studies can be referred to as "replications" while differing widely in terms of participants, time, place, or methods of data-gathering and analysis.

We do not want to add to the confusion here with more definitions! Rather, we intend to present you with three existing definitions and define a practical use for each within a systematic, cumulative approach to replication work. The intention is to set out firm replication research series which are interdependent, and in which you can participate – first through *close*, followed by *approximate*, and perhaps then supplemented by *conceptual* replications.

Having said this, let us begin by discarding for our present purposes the "classic" sense of replication. "Exact", "literal", or "direct" replications – essentially doing the same study again – and wherein the manipulation, measurement of variables, and procedures are all kept the same, are just not possible in social sciences research. Indeed, in the strictest sense, neither are they possible in the so-called "hard" sciences – the size of a test tube may vary, for example, as the quality and texture of a chemical may differ from one lab to another. A replicated study can never really be the same, be it repeated by the same researcher in the same context, or others. Time intervals play their part, as do participant variables, changed contexts, and a host of other variables. As Rosenthal observed[3] (1991) "[In behavioral research] . . . replications are only possible in a relative sense". The best we can strive for is to monitor and control for the conditions that might otherwise affect outcomes.

Thus, if exactness is accepted as an unreachable objective, we will need to make judgments as to just how close we need to get to permit those sound comparisons to be made between the original and replicated studies.

5.1.2.1 Close Replication

We have already prepared the way for close replication by means of the routine checking and critique we encouraged in the previous chapter. By having got "up

close and personal" with your chosen target study you are now acquainted both with the methods and procedures used in that study and with that author's *modus operandi* – specifically the large number of decisions he or she would have had to make along the way.

Armed with this information, you should now be in a position to understand, or conjecture, how the results you were presented with might have come about. All this should mean you are also a little more aware of how you will be able to further the knowledge gained from those outcomes. Ideally, you would even have made those notes in the margins during your critique, identifying new aspects or questions the original author did not think about or take into account in his or her execution of the research, and which you felt might have affected the outcomes.

One of the principal reasons for conducting a replication is to increase the confirmatory power of the original study. Replications proceed by degrees: with every pertinent modification in subsequent replications the confirmatory power of the study may increase and, potentially, be generalized to further or wider applications. However, as with replication in the pure sciences, we must proceed cautiously and at a deliberate pace, be systematic and careful in the application and interpretation of each modification. Just as this cautiousness can eventually lead us to an increase in confirmatory power, changes in several types of variables *simultaneously* (in one study, for example) take us further away from the original study and means, for example, that failure to replicate the original outcome becomes much more difficult to interpret or pin down to a specific variable.

All this argues for an initial, cumulative, pre-planned series of replication attempts of a target study – a replication scheme if you will – whereby *only one* major variable is modified each time (it can be added *or* removed, of course!) and all others are kept as constant as possible to be better able to single out the kind of influences each has on the dependent variable.

Such a systematic series of progressive changes we will call a set of *close replications* such as those you see outlined in the next section. Each would ideally be carried out by a number of independent "labs" or researchers who would present their data before moving on the second, and so on. To do so efficiently requires the integration of these teams into a systematic program.

Such a progression also implies the need for some kind of external executive control or planning regarding what is replicated, how, and when. While this process is more measured and deliberate in its approach to research data-gathering than you might be used to, it better enables us methodically to build up sufficient comparative evidence about a study's validity or generalizability and involve large groups of like-minded researchers working in different contexts toward a common aim.

Sample Series of Close Replications

Read the following study – F. Pichette, L. de Serres, and M. Lafontaine (2012). "Sentence reading and writing for second language vocabulary acquisition."

Applied Linguistics, 33.1, 66–82. Then *re*read the three sections marked "Research questions", "Participants", and "Procedure", paying particular attention to the description of the variables and the reasons why the study was carried out.

Below is a series of imagined plans for close replications of the study, each taking up a different key aspect of the original (e.g., participants, time, and task condition). For each sample series, you are presented with:

a. the **variable modification envisaged**;
b. a **brief justification for the study** based on what the original (target) study has presented or reported; and
c. a **statement with an imaginary outcome/conclusion** which attempts to suggest how further close (or approximate) replications might seek to fine tune these outcomes.

PARTICIPANTS: STATUS SERIES

VARIABLE MODIFICATION: L2 STATUS (ESL vs EFL)

JUSTIFICATION: Studies of native Spanish L2 English students in Spain studying EAP (Author a, 1998, Author b *et al.*, 2015) have shown that advanced-level EFL **university students (18–22)** also acquire new words better through writing. Would *EFL* students of similar proficiency show similar effects to those found in the original study?

ORIGINAL STUDY: "203 French-speaking **ESL** students enrolled in . . . university. 18–53. Mean age 24.2." Course being studied is unclear.

REPLICATION: Attempt to replicate research question 1 with a comparable number and status of **EFL** university students and with a similar mean age.

OUTCOME: Similar positive effect – average score lower – but less pronounced than the original – for words read compared with words written. Tests show only slight drop between immediate and delayed recall tests for both concrete and abstract words – unlike original study. Probably worth investigating EFL students with other L1s. Consider also outcomes for an approximate replication study with L2 status and immediate/delayed recall variables?

****

VARIABLE MODIFICATION: LEARNING CONTEXT (HIGH SCHOOL vs UNIVERSITY)

JUSTIFICATION: Previous recall studies carried out in high schools (Author a, 2014a, Author a *et al.*, 2014b) indicate L2 recall of concrete items

varies along with age (14–18). Would ESL *high school* students show similar recall as in the original study?

ORIGINAL STUDY: "203 French-speaking ESL students enrolled in . . . **university**. 18–53. **Mean age 24.2**." Wide age range. Results not broken down in terms of ages.

REPLICATION: Attempt to replicate research question 1 with a comparable number and status of ESL high school students focusing on a narrower age range.

OUTCOME: Similar effect as in the original found here in ESL high school samples. However, *F*-tests showed similarly high recall for the second (delayed) recall – in contrast to the original study. Narrower age range across high school years sampled may indicate younger people are somehow better able to recall???

****

VARIABLE MODIFICATION: DEGREE STATUS (ESL students reading Literature/vs??)

JUSTIFICATION: Author a, 2001, Author b, 2006, and Author c, 2007 demonstrated that ESL students studying ESL Business Studies and English Language majors in the US and Canada had statistically different recall for L2 English pseudo-words through writing. Would outcomes from the original study be replicated across participants from differing degree areas?

ORIGINAL STUDY: Unspecified degree status. Check with authors to see if large numbers of participants indicated *various* degree affiliations involved.

REPLICATION: Attempt to replicate research questions 1 and 2 with a comparable number and status of ESL university students focusing on different degrees being read.

OUTCOME: ESL Business Studies degree students and English Language degree students show similar recall to that found in the original across at initial time sample but the former degree students reveal considerably greater losses at the second sample. Consider replications with other degrees?

****

VARIABLE MODIFICATION: L1 CHINESE vs L1 FRENCH

JUSTIFICATION: Studies of native Chinese ESL students (Author a, 1998, Author b *et al.*, 2015) showed that these intermediate and advanced-level ESL university students (18–22) acquire new words better through writing and retain a high level of recall across a number of weeks.

ORIGINAL STUDY: "**French-speaking** ESL students enrolled in . . . university."

REPLICATION: Attempt to replicate research questions 1 and 2 with a comparable number of Chinese-speaking ESL university students.

OUTCOME: Stronger effect than found in the original, particularly in the second, delayed sample. Would other L1s reveal similar outcomes? Is it L1 or the way the L1 students approach the task?

TIME: DATA-GATHERING SERIES

VARIABLE MODIFICATION: IMMEDIATE/DELAYED RECALL

JUSTIFICATION: Many studies of L2 vocabulary recall (e.g., Author a, 1999, Author b, 2013, Author c *et al.*, 2016) with intermediate and advanced learners show that positive recall weakens considerably over periods longer than 1 month. It is pedagogically useful to find a longer-lasting effect than in the original study. Original study only took two samples of recall, one immediate ("surprise") and one delayed after a week.

ORIGINAL STUDY: "Delayed recall suggests this superior recall for writing tasks over reading and concrete over abstract words disappears over time . . . With a more longitudinal version of this study, it would be interesting to see if this levelling pattern would continue over a longer period . . . if superiority for writing would continue to degrade up to the point of disappearing completely . . . [and] there remains the ever-present question of the extent to which recall measures actually measure acquisition."

REPLICATION: Attempt to replicate research question 3 with similar participant numbers and nature but with four samples taken over the original week period ("surprise") and an additional two (announced beforehand) over a further one-month period. Perhaps recall is enhanced through the number of samples taken and/or the unannounced/announced nature of these?

OUTCOME: Beneficial effect originally noted for superior recall for writing tasks over reading and concrete over abstract words is confirmed to diminish only slightly over the four initial samples. However, the rate of loss observed in the later samples was also not as pronounced. Possible idea for approximate replication to replicate original with an altered writing condition (see below) and delayed recall tests over longer periods??? Was the increase in number of tests (increased to four) an important influence on the effect noted? Worth replicating with more frequent tests over same time period?

TASK CONDITION SERIES

VARIABLE MODIFICATION: ALTERED WRITING TASK

JUSTIFICATION: Original study (*Limitations*) suggests a variety of writing tasks might shed more light on recall effects.

ORIGINAL STUDY: "For the writing task, participants were to write three sentences per item, with each sentence containing the target L2 word. The target word, accompanied by its definition in the L1, preceded each sequence of three sentences."

REPLICATION: Attempt to replicate research question 1 with a comparable number and status of ESL university students and comparing effects on recall, adding a fill-in-the-blanks exercise with sentences already provided rather than invented by participants.

OUTCOME: Relatively small, but statistically significant, difference seen between the two writing conditions. Useful to have a further close replication with other writing conditions? Consider approximate replication with altered writing condition *and* L2 status?

» Activity 13a

Look at the final possible "outcomes" in the replication series above. Can you suggest – given these hypothetical outcomes of the replication undertaken – what a *further* close replication study might focus on?

» Activity 13b

Choose ONE section from the complete paper above (F. Pichette, L. de Serres, and M. Lafontaine (2012). "Sentence reading and writing for second language vocabulary acquisition." *Applied Linguistics*, 33.1, 66–82) from *Participants*, *Items*, *Tasks*, *Time allotted for completing the tasks*, *Procedure, Scoring and analysis,* or *Limitations*. Think about TWO further possible series involving a *minimal* variable modification within that section. Then make similar notes to those above describing how a subsequent close replication with that variable change might shed new light on the effects observed in the original study.

Our encouraging you to start with close replication serves two purposes: first, it acknowledges that no replications in the field of social science can be truly "exact" and that this means any attempt to revisit a previous study's findings has perforce to accept some, ideally minimal, modification. Second, such a minimal change allows us to make a relatively "safe" comparison between our outcomes and those of the original and thereby feed in to the body of knowledge arising from the target study and, now, its subsequent replications.

5.1.2.2 Approximate Replications

We are presenting approximate replications as the *next* (rather than an alternative) step *after* close replications since we regard them as a logical *consequence* of previous close replications – rather than a starting point for investigations. For us, these will logically follow on from one another – an approximate replication thereby ideally taking up the issues raised in the previous series of close replication outcomes.

Again, comparative evidence will be the key throughout: we will eventually want to be able to make a fair comparison between what comes out of our approximate replication and the target study. And to do so, we will also want to make sure as many of the elements as possible of the original study are kept constant, while our focus will now be on the effects of *two* variables on outcomes.

Many published replications you may read in AL are labelled "approximate", "partial", or "conceptual" in nature; however, more often than not these will see modifications in two *or more* elements, crucial or otherwise, to the original study.

As you read in the previous sections, each modification inevitably takes you just that bit further away from the original study and so makes the essential comparisons all that more problematical. Indeed, many of these "approximate" replications actually represent initial replications of the target study (i.e., they do not derive from a previous *close* replication attempt); under these circumstances the reader will need to approach the findings with care and be informed of the effect of each variable – and all of them.

» Activity 14

Below are a number of studies which are described or framed as "approximate" or "partial" in nature but which have not been consequent to a previous close replication.

Locate the studies and focus on the *Procedures, Results,* and *Discussion* section in each to

i) isolate the variables which appear to have been modified from the original/target study and those that have been held constant; and then

ii) hypothesize about what consequences there might be for the interpretation of outcomes of the replication as a result of this combination of variable modifications.

1. Roller, C. & Matambo, A. (1992). Bilingual readers' use of background knowledge in leaning from text. *TESOL Quarterly*, 26, 129–141.
2. Mullock, B. (2006). The pedagogical knowledge base of four TESOL teachers. *Modern Language Journal*, 90, 48–66.
3. Schauer, G. (2006). Pragmatic awareness in ESL and EFL contexts: Contrast and development. *Language Learning*, 56, 269–318.

We will now return to the Sample Series of Close Replications outcomes above and see if/how an ensuing *approximate* replication might provide useful further data.

For example, in the PARTICIPANTS: STATUS SERIES above, the variable modification carried out (L2 STATUS (ESL vs EFL)) revealed a similar, if weaker, positive effect than the original with the EFL participants. Interestingly, and unlike the original study, "tests show[ed] only [a] slight drop between [the] immediate and delayed recall tests for both concrete and abstract words". As we made a note there, it might be of interest now to look at two of our previously studied variables in an approximate replication (L2 status and immediate/delayed recall) to see how far the somewhat better recall data observed above might be replicated with the increased sampling undertaken in the TIME: DATA-GATHERING SERIES. That close replication revealed a similar outcome to the original and ". . . the rate of loss observed in the later samples was also not as pronounced".

Thus, we might write out the *approximate* replication attempt as:

VARIABLE MODIFICATION: L2 STATUS (EFL) & IMMEDIATE/DELAYED RECALL

JUSTIFICATION: Previous close replications with EFL participants (rather than ESL) revealed a positive, albeit weaker, effect than the original with EFL participants. Modifications to the recall time sampling also showed similar results to the original but with less pronounced loss noted in the later samples.

ORIGINAL STUDY: "203 French-speaking **ESL** students enrolled in . . . university. 18–53. Mean age 24.2."

"Delayed recall suggests this superior recall for writing tasks over reading and concrete over abstract words disappears over time . . . With a more longitudinal version of this study, it would be interesting to see if this levelling pattern would continue over a longer period . . . if superiority for writing would continue to degrade up to the point of disappearing completely . . .

[and] there remains the ever-present question of the extent to which recall measures actually measure acquisition."

APPROXIMATE REPLICATION: Attempt to replicate research question 3 with EFL participants and with four samples taken over the original week period ("surprise") and an additional two ("announced beforehand") over a further one-month period.

Perhaps recall is seen to improve more for the EFL participants with the modifications to the time sampling.

OUTCOME: Inconclusive outcomes. Some of the participants showed less pronounced fall in recall while others stayed at much the same level. Recommendations to attempt to replicate study across a number of other EFL populations.

As you see in the hypothetical study above, the results of a replication may well point to the need for its own replication before moving on in the series. While the outcome may have us wondering about the usefulness of a further replication with another variable – such as the "altered writing condition" variable modification above – we first need to gather enough data from enough replications of the two variables at hand to begin to form a picture of what the interaction might be telling us about the original study and its generalizability.

Hopefully, you will also notice the piecemeal nature of the research process through replication. As we mentioned in the Introduction, your objective in this approach to research is not to move from one context or variable or instrument or procedure to another and thereby accumulate as much data as possible using different forms of data collection or from a number of disparate contexts and participants. This would be more characteristic of what we do in follow-up studies or extensions of the target study. Replication rather rewards the patient, methodical, cumulative approach to data-gathering from individual researchers or research teams working on the same study.

» Activity 15

You are going to design some approximate replications. Look back at the close replication series on pp. 73–77, and your own answers in Activities 13a and 13b. Think carefully first about what might justify the study of two variables you now wish to pursue in each replication of the original study. Discuss with someone what you might expect as a possible outcome of a replication involving both variables. Then devise a set of THREE approximate replications in the way you saw above and present them in note form as you saw on these pages, together with some hypothetical outcomes.

» Activity 16

Here are some similar notes on FOUR hypothetical *close* replications of the Bitchener and Knoch (2010) key study (see Introduction, p. 10)

Imagine that you now wanted to move on to some approximate replications of this same study, based on the outcomes from the close replications you read about below. Focus on what has been undertaken in each close replication and complete similar notes as those above for the approximate replications to show how two of the variables studied might now be usefully focused on. Justify why any possible outcomes you suggest might be of interest to the field.

1

VARIABLE MODIFICATION: PARTICIPANT STATUS/ORIGIN

JUSTIFICATION: Original study indicated a non-specified array of participants. Previous studies by Author, 2010 and Author, 2012 amongst others show that Chinese ESL university students did not benefit from written corrective feedback as much as non-Chinese ESL participants and suggested that previous school experience might have conditioned the former to responding to a different type of written feedback on their work. Will delimiting the participants to individual countries or educational background reveal more detailed information about the extent of the beneficial effects on accuracy?

ORIGINAL STUDY: "Most of the participants were from a range of East and South Asian countries . . ."

REPLICATION: Attempt to replicate original research questions 1 and 2 with a comparable number and proficiency status (i.e., "advanced") of ESL university students from one country/educational background.

OUTCOME: Groups receiving direct feedback outperformed the control in the immediate post-test. Indirect feedback not as successful. Over the ten-week period improvement in the direct groups weakened considerably.

2

VARIABLE MODIFICATION: TIME SPENT IN US EDUCATIONAL SYSTEM

JUSTIFICATION: Communication with authors confirms participants varied greatly (range = zero through eight years) regarding the amount of their previous education having been undertaken in the US. Several previous studies with ESL students in a similar context indicated that

(continued)

(continued)

longer experience of typical direct written feedback procedures at high school might increase accuracy performance at university.

Will selecting treatment and control groups on the basis of time spent in the US educational system present similar or distinct effects on accuracy?

ORIGINAL STUDY: No reference to variable – confirmed by the authors.

REPLICATION: Attempt to replicate both research questions with a comparable number and proficiency status (i.e., "advanced") of ESL university students.

OUTCOME: Groups receiving both direct and indirect feedback and with more than five previous years in US education system outperformed the control in the immediate post-test and over the ten-week period. Groups with less than two years' experience performed less accurately both in treatment and control groups with the indirect feedback noticeably weaker over the ten-week period than the other treatment groups.

3

VARIABLE MODIFICATION: MOTIVATION

JUSTIFICATION: Studies looking at the effect of instrumental motivation (e.g., Author 2001, 2007 and Author *et al.* 2010) have shown that the L2 productive skills (speaking and writing) appear to benefit in terms of accuracy and error detection when such motivation is identified in ESL and EFL participants. Will groups of instrumentally-motivated ESL students show similar outcomes with this treatment of corrective feedback?

ORIGINAL STUDY: No specific reference to motivational differences within or across groups. Authors confirm this was not a consideration in selection. Anecdotal evidence presented from some participants: "high degree of motivation to improve accuracy . . . assumed . . . evident from the comments made by many of the students . . ."

REPLICATION: Attempt to replicate both research questions with a comparable number and proficiency status (i.e., "advanced") of ESL university students.

OUTCOME: Groups receiving both direct and indirect feedback and identified with high instrumental motivation indices outperformed the control in the immediate post-test and over the ten-week period. Very similar, but not significantly better, results to the original study on both research questions. Indirect feedback group tended to maintain the initial superior longitudinal effect on accuracy over ten-week testing period although – as in the original study – not as strong as the direct feedback group.

4

VARIABLE MODIFICATION: L2 PROFICIENCY LEVEL

JUSTIFICATION: Author *et al.* (1998, 2000, and 2008) produced a series of studies contrasting "upper-intermediate and advanced" German and Chinese L2 English writers at two universities in England. Participants received *indirect* corrective feedback on their writing through three different instruments (see above). Will selecting treatment and control groups on the basis of L2 proficiency level ("upper-intermediate") present similar or distinct effects on accuracy?

ORIGINAL STUDY: "63 advanced level students". Similar method of operationalizing "upper-intermediate" levels used in this replication using published proficiency level tests. The authors suggest in the "Discussion and conclusions" that "further research is needed to clarify the relative effectiveness of all types of direct and indirect feedback . . . *given to L2 writers of different proficiency levels*" (authors' emphasis). They also note that previous work by the same authors "with lower proficiency writers" found *no* difference between groups who received different types of *direct* feedback.

REPLICATION: Attempt to replicate both research questions with a comparable number of ESL university students and with proficiency status classified as "upper-intermediate".

OUTCOME: Similar results to the original study. Groups receiving direct feedback outperformed the control in the immediate post-test. Indirect feedback not as successful. Over the ten-week period, improvement in the direct feedback groups remained strong.

5.1.2.3 Conceptual Replications

The preceding section focused on replications that took their cue from previous close replication findings in the specific study of interest. As we move further along our hypothetical continuum of research which focuses on comparative outcomes between studies, we inevitably arrive at a greater distance from the target study than in the previous cases.

Our starting point is no longer what previous close or approximate replications have discovered – although such findings might well feed into how we design the study. A conceptual replication would see even less of the target study being replicated; our objective as researchers requires a different focus. Now we will focus on the findings from the target study but not with the objective of direct comparison with our own replication findings. We will rather attempt

to widen the application, relevance, or generalizability of the underlying theory or hypotheses of the original study by, for example, using a completely different database and related collection procedure (e.g., observation against self-report, qualitative supplementing original quantitative data), operational definition or methodology, or method of analysis.

There are a number of reasons why we might want to do this. First, you might come across a study which was undertaken some considerable time ago but you see continues to be cited in the literature and whose findings are still reported as crucial in an area of your interest. If findings are understood to remain particularly significant for the field – despite their apparently dated nature – it might be a cue for you to revisit them with more recent analytical methods or in the light of advances in data collection or methodology. Such an approach might be a valuable contribution if your aim was thereby to reinforce validity or reliability, for example. The consequence of your study would not necessarily *replace* the original study – albeit you would have doubtless tried to replicate it as closely as possible with the new methodology or analytical method. Rather, the two studies would stand as evidence in their respective times and with their corresponding methods and outcomes.

» Activity 17

Research three or four "historical" studies from your own area of interest and which continue to be frequently cited in the literature as remaining significant for that area. For each study

a. highlight its perceived importance for current debate;
b. summarize the main findings which remain significant;
c. isolate the methodology used to collect the data;
d. identify the method(s) used to analyze the data obtained;
e. suggest if any more recent methodological approaches might be useful in a replication update;
f. consider how alternative, more recent, data analysis approaches might replace or supplement that originally collected.

Apart from using a conceptual replication to update a target study, we may want to look at the underlying premise or theory of the target study itself with a view to providing more information on its application or generalizability. We will not be focusing on the original study's reliability or validity, but rather on the extent to which the outcomes might be generalized to other constructs, or methodological or analytical approaches. We might also return to the target study to provide further tests of a hypothesis using different experimental approaches.

Our aim is to examine the same underlying theory as the original. In this way, the use of replication extends beyond confirmation of the original findings toward theory and model building.

However, and unlike in the so-called "hard" or pure sciences, our operationalization of AL constructs will need to be adapted to different learning contexts, or methodologies, or time periods, and so on. Often, what is defined in one context in the original study may not square with our own. For example, the construct "listening comprehension" can be operationally defined in various ways and apply in various contexts. We need, therefore, to see how far what has been revealed about listening comprehension in one context applies equally in another. Conceptual replication encourages us to test multiple manifestations of the construct and thereby try to arrive at a consensus about the generalizability of the outcomes.

If the theory, or hypothesis, or treatment is seen to be supported across several different constructs, or methodologies, or analysis methods, we can begin to have confidence in that theory, beyond the effects observed in the one, original, study.

As with our updating through replication, we will be employing different procedures from the original study. These might be in our definition of the principal construct at hand, the relationship between constructs, the way data is gathered, and so on. Imagine, for example, that Researcher X (1998) carried out a close replication of Researchers Y and Z's (1993) study into the learning strategies L2 English learners used when going about learning new vocabulary. Both studies used retrospective data collection methods albeit Researcher X chose to modify the original using similar proficiency level L2 German students.

Both papers reported similar findings in terms of the main metacognitive strategies reported but there were large differences reported regarding the individual strategies used for recording and memorizing new vocabulary. This inconsistency surprised both Researcher X and a number of others in the field given the findings of similar studies carried out since Y and Z's seminal work. However, Researcher X's experimental procedures diverged somewhat from Y and Z in the individual retrospective interviews. The original paper had participants identifying individual strategies used from a checklist while in X's replication participants were asked to recall as many strategies as they considered important. X decided that outcomes might have been affected by the fact that the original procedure restricted responses to a predesigned list of options contrasting with the more open nature of X's own data collection. Moreover, other researchers had criticized X's method of teaching the control group about learning strategies before the study itself. They argued that the resulting priming effects on this group meant that participants were already aware of the kind of strategies to look out for in their subsequent reporting of events.

As a result of these concerns, Researchers A and B decided to replicate Y and Z's study conceptually by developing and testing a more comprehensive model of procedures for identifying learning strategies, streamlining the original procedure

and introducing think-aloud protocols alongside retrospective interviews as additional sources of data.

This example illustrates how one study can feed into another through a conceptual replication (for examples of this process for partial replications, see also Eckerth, 2009 and McManus and Marsden, 2018 in Chapters 6 and 7). The data obtained can be further enriched by modifying or building upon the methodology used in the original (or subsequent close replications of it). The result is, again, cumulative knowledge – rather than the mere accumulation of data from one-off studies – which can be collated and analyzed with a view to further theory testing and building.

» Activity 18

Below are three abstracts from studies which are described as, or understood to be, "conceptual replications" of previous work. Read the extracts carefully, underlining where the author described the changes made to the original study in his or her replication. Explain what characteristic(s) described makes this a conceptual replication. If you think you need more information, read the full article before you make your decision.

Paper 1

THE EFFECTS OF INPUT ENHANCEMENT ON GRAMMAR LEARNING AND COMPREHENSION: A MODIFIED REPLICATION OF LEE (2007) WITH EYE-MOVEMENT DATA

www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/effects-of-input-enhancement-on-grammar-learning-and-comprehension/FA73F01ADB6A7B4148AD25D697F401D7.

PAULA WINKE (*Studies in Second Language Acquisition*, 35.2, 323–352).

In his 2007 study "Effects of Textual Enhancement and Topic Familiarity on Korean EFL Students' Reading Comprehension and Learning of Passive Form," Lee demonstrated that learners were better able to correct written sentences that contained incorrect English passive forms after exposure to texts flooded with enhanced (versus nonenhanced)[4] passive forms. But with enhanced forms, learners did worse on comprehension tests, which arguably demonstrated a trade-off: More attention to forms resulted in less to meaning. In this study, a conceptual replication of Lee's using eye-movement data, I assessed how English passive construction enhancement affects English language learners' (a) learning of the form (via pre- and posttest gains on passive construction tests) and (b) text comprehension. In contrast to Lee's results, I found enhancement did not significantly

increase form correction gain scores, nor did enhancement significantly detract from comprehension. There was no trade-off effect. Form learning and comprehension did not correlate. By recording learners' eye movements while reading, I found enhancement significantly impacted learners' noticing of the passive forms through longer gaze durations and rereading times. Thus, enhancement in this study functioned as intuitively and originally (Sharwood Smith, 1991, 1993) proposed; it promoted noticing, but, in this case, without further explicit instruction, it appeared to have done little else.

Paper 2

DISCOURSE PROCESSING EFFORT AND PERCEPTIONS OF COMPREHENSIBILITY IN NONNATIVE DISCOURSE: THE EFFECT OF ORDERING AND INTERPRETIVE CUES REVISITED

www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/discourse-processing-effort-and-perceptions-of-comprehensibility-in-nonnative-discourse/1AA1DA0EA88AC7060312852465DCC5A5.

ANDREA TYLER AND JOHN BRO (*Studies in Second Language Acquisition*, 15.4, 505–522).

The study reported here extends Tyler and Bro's (1992) investigation of the sources of native speakers' perceptions of incoherence in English text produced by nonnative speakers. Using paper-and-pencil tasks, the original study examined two competing hypotheses: (a) The primary source of interference was the order in which the ideas were presented versus (b) the primary source of interference was mismatches in discourse structuring cues. They found no effect for order of ideas but a strong effect of discourse structuring cues. In the present study, 80 subjects were tested on the same texts as those used in Tyler and Bro (1992) but using microcomputers. Subjects rated the text for comprehensibility and answered three questions concerning the propositional content. The computer format represented a more sensitive measure of subjects' reactions to the text because it did not allow looking back and because it provided information concerning differences in reading time for each manipulation. Once again, the results of the comprehensibility ratings showed a strong effect for miscues and no significant effect for order of ideas. Results of the true/false questions indicated that presence of miscues affected subjects' comprehension of the propositional content but that order of ideas had no discernible effect. Finally, reading time results also showed a strong effect for miscues and a mixed effect for order of ideas, suggesting that order of ideas does make a minor contribution to comprehensibility.

(continued)

(continued)

Paper 3

ON THE SECOND LANGUAGE ACQUISITION OF SPANISH REFLEXIVE PASSIVES AND REFLEXIVE IMPERSONALS BY FRENCH- AND ENGLISH-SPEAKING ADULTS

http://journals.sagepub.com/doi/abs/10.1191/0267658306sr260oa.
ANNE TREMBLAY (*Second Language Research*, 22.1, 30–63).

This study, a partial replication of Bruhn de Garavito (1999a; 1999b), investigates the second language (L2) acquisition of Spanish reflexive passives and reflexive impersonals by French- and English-speaking adults at an advanced level of proficiency. The L2 acquisition of Spanish reflexive passives and reflexive impersonals by native French and English speakers instantiates a potential learnability problem, because (1) the constructions are superficially very similar (*se* V DP) but display distinct idiosyncratic morphological and syntactic behaviour; (2) neither exists in English, and the reflexive impersonal does not exist in French; and (3) differences between the two are typically not subject to explicit instruction. Participants – 13 English, 16 French and 27 Spanish speakers (controls) – completed a 64-item grammaticality-judgement task. Results show that L2 learners could in general differentiate grammatical from ungrammatical items, but they performed significantly differently from the control group on most sentence types. A look at the participants' accuracy rates indicates that few L2 learners performed accurately on most sentence types. Grammatical and ungrammatical test items involving [+animate] DPs preceded or not by the object-marking preposition *a* were particularly problematic, as L2 learners judged them both as grammatical. These results confirm that the L2 acquisition of Spanish reflexive passives and reflexive impersonals by French- and English-speaking adults instantiates a learnability problem, not yet overcome at an advanced level of proficiency.

» Activity 19

You are now presented with three abstracts from recent studies which we will imagine have been singled out as *in need of conceptual replication*. Look up the studies themselves if you need more details about the background, procedures and methodology.

a) Read the details provided about each study; and
b) pay particular attention to constructs/operational definitions, data sources, procedures, or analysis methods used and justify how you might

go about a *conceptual replication* study which could conceivably add useful information to further build theory or help generalize the results.

Paper A

AFFECT TRUMPS AGE: A PERSON-IN-CONTEXT RELATIONAL VIEW OF AGE AND MOTIVATION IN SLA

http://journals.sagepub.com/doi/abs/10.1177/0267658315624476?journalCode=slrb.

SIMONE E. PFENNINGER AND DAVID SINGLETON (*Second Language Research*, 32.3, 311–345).

Recent findings indicate that age of onset is not a strong determinant of instructed foreign language (FL) learners' achievement and that age is intricately connected with social and psychological factors shaping the learner's overall FL experience. The present study, accordingly, takes a participant-active approach by examining and comparing second language (L2) data, motivation questionnaire data, and language experience essays collected from a cohort of 200 Swiss learners of English as a foreign language (EFL) at the beginning and end of secondary school. These were used to analyze (1) whether in the long run early instructed FL learners in Switzerland outperform late instructed FL learners, and if so the extent to which motivation can explain this phenomenon, (2) the development of FL motivation and attitudes as students ascend the educational ladder, (3) the degree to which school-level variables affect age-related differences, and (4) learners' beliefs about the age factor. We set out to combine large-scale quantitative methods (multilevel analyses) with individual-level qualitative data. While the results reveal clear differences with respect to rate of acquisition in favor of the late starters, whose motivation is more strongly goal- and future-focused at the first measurement, there is no main effect for starting age at the end of mandatory school time. Qualitative analyses of language experience essays offer insights into early and late starters' L2 learning experience over the course of secondary school, capturing the multi-faceted complexity of the role played by starting age.

Paper B

THE INFLUENCE OF FOREIGN SCRIPTS ON THE ACQUISITION OF A SECOND LANGUAGE PHONOLOGICAL CONTRAST

http://journals.sagepub.com/doi/abs/10.1177/0267658315601882.

LIONEL MATHIEU (*Second Language Research*, 32. 2, 145–170).

Recent studies in the acquisition of a second language (L2) phonology have revealed that orthography can influence the way in which L2 learners

(continued)

(continued)

come to establish target-like lexical representations. Most of these studies, however, involve language pairs relying on Roman-based scripts. In comparison, the influence of a foreign or unfamiliar written representation on L2 phonological acquisition remains understudied. The present study therefore considers the effects of three L2 scripts on the early acquisition of an Arabic consonantal contrast word-initially (e.g. /ħal/–/χal/). Monolingual native speakers of English with no prior knowledge of Arabic participated in a word-learning experiment where they were instructed to learn six pairs of minimally contrastive words, each associated with a unique visual referent. Participants were assigned to one of four learning conditions: no orthography, Arabic script, Cyrillic script, and Roman/Cyrillic blended script. After an initial learning phase, participants were then tested on their phonological knowledge of these L2 minimal pairs.

The results show that the degree of script unfamiliarity does not in itself seem to significantly affect the successful acquisition of this particular phonological contrast. However, the presence of certain foreign scripts in the course of phonological acquisition can yield significantly different learning outcomes in comparison to having no orthographic representation available. Specifically, the Arabic script exerted an inhibitory effect on L2 phonological acquisition, while the Cyrillic and Roman/Cyrillic blended scripts exercised differential inhibitory effects based on whether grapheme–phoneme correspondences activated first language (L1) phonological units. Besides revealing, for the first time, that foreign written input can significantly hinder learners' ability to reliably encode an L2 phonological contrast, this study also provides further evidence for the irrepressible hold of native orthographic rules on L2 phonological acquisition.

Paper C

INPUT PROCESSING AT FIRST EXPOSURE TO A SIGN LANGUAGE

http://journals.sagepub.com/doi/abs/10.1177/0267658315576822.

GERARDO ORTEGA AND GARY MORGAN (*Second Language Research*, 31.4, 443–463).

There is growing interest in learners' cognitive capacities to process a second language (L2) at first exposure to the target language. Evidence suggests that L2 learners are capable of processing novel words by exploiting phonological information from their first language (L1). Hearing adult learners of a sign language, however, cannot fall back on their L1 to process novel signs because the modality differences between speech (aural–oral) and sign (visual-manual) do not allow for direct cross-linguistic influence. Sign language learners might use alternative strategies to process input expressed in the manual channel. Learners may rely on iconicity, the direct relationship

between a sign and its referent. Evidence up to now has shown that iconicity facilitates learning in non-signers, but it is unclear whether it also facilitates sign production. In order to fill this gap, the present study investigated how iconicity influenced articulation of the phonological components of signs. In Study 1, hearing non-signers viewed a set of iconic and arbitrary signs along with their English translations and repeated the signs as accurately as possible immediately after. The results show that participants imitated iconic signs significantly less accurately than arbitrary signs. In Study 2, a second group of hearing non-signers imitated the same set of signs but without the accompanying English translations. The same lower accuracy for iconic signs was observed. We argue that learners rely on iconicity to process manual input because it brings familiarity to the target (sign) language. However, this reliance comes at a cost as it leads to a more superficial processing of the signs' full phonetic form. The present findings add to our understanding of learners' cognitive capacities at first exposure to a signed L2, and raises new theoretical questions in the field of second language acquisition.

At this point, it is worth remembering our caveats at the beginning of this section. In executing a conceptual replication, our aim is very different from the previously-seen close and approximate types. Since we are not modifying variables with an aim of validating the original study, the outcome of a conceptual replication cannot be construed as an indisputable test of the original study's validity or reliability. So, for example, imagine our target study tested the hypothesis that listening comprehension can be improved with an intervention based on oral responses. This was tested successfully with a group of teenage L2 English learners in France; however, a subsequent conceptual replication which modified the response mode to written as well as redefining listening comprehension operationally and with a group of adult L2 German learners in the US was not successful. The original study's outcomes are not now necessarily to be questioned. There might be myriad reasons for the different outcome. However, useful further information might still have been gleaned concerning the underlying hypothesis of the original and even the extent to which listening comprehension ability is susceptible to successful intervention at all.

Thus, by extending the domain of the original research we can still obtain potentially useful information *through*, if not *about*, that study. The procedure is designed to test not whether a particular outcome holds across small modifications in variables, but rather across larger concepts such as the operationalizations of constructs. As we said above, this is particularly useful in a field such as AL where it is easy to find ambiguity and difficult to find a consensus as to their acceptable definition.

You may recall here one of the key papers – Pichette *et al.* (2012).[5] There, in the abstract, we read that the authors were looking at ". . . the relative

effectiveness of reading and writing sentences for the incidental acquisition of new vocabulary in a second language". To do so, they present three research questions (p. 70) with a number of constructs – some of which we have put in italics:

> Question #1
>
> For intermediate and advanced L2 students, does sentence writing lead to higher *vocabulary gains* relative to sentence reading?
>
> Question #2
>
> For intermediate and advanced L2 students, does *recall* vary according to the concreteness of target words?
>
> Question #3
>
> Does the impact of task and *concreteness* change over time?

Your close reading and critical address (see Chapter 2) of the text should have encouraged you to look out for the operationalization of these constructs within the study. At that point such attention to how constructs were defined in the study was a question of whether we found them acceptable as a practical, testable, observable, and/or measurable quality within that study. Now we need to think about how far that operationalization might be modified to permit useful conceptual replicas to be undertaken.

So, for example, in the above study, reading further ahead we discover that "recall" was defined as follows (p. 71):

> The recall task chosen was cued recall, which requires the participants to provide the L2 word via a clue offered by the experimenter. The measured knowledge is thus of a productive, not receptive, nature. Cued recall is recognized as sensitive to word forms, since the person tested does not have to recognize the L2 form, but retrieve it from memory and produce it correctly. Since the experiment included abstract words, the clues were L1 French definitions, since the use of illustrations would be difficult, if not impossible.

Later on we read that this operational definition was mooted to have had its drawbacks (p. 78–79):

> . . . . many students had probably guessed that some sort of recall test would be given as a follow-up task, given the fact that they are frequently solicited for participating in studies during their degree program. And there remains the ever-present question of the extent to which recall measures actually reflect acquisition.

It appears that a conceptual replication might usefully test a different definition of this construct, perhaps one that modifies the kind or amount of cue provided. For example, you might have read in a related study that recall of vocabulary items had been tested in a less obvious, or less typical, or more informal way such as by producing a short, oral text. You might even feel there

was useful information to be gleaned by testing recall here receptively rather than productively, as in the original.

Meanwhile, "vocabulary gains" are not clearly defined. "Gains" would seem to imply some kind of improvement from a former measured state. However, as we read here, the measurement seems to be looking at the relative difference between reading and writing as a better aid to recall. Thus, as the researchers themselves point out in their "Limitations", ". . . there remains the ever-present question of the extent to which recall measures actually reflect acquisition". As we mentioned earlier when discussing this paper for close and approximate replication (see. p. 73), we could argue that an operational definition of "gain" which covered such a short period is unlikely to present enough information for us to measure the extent to which the new vocabulary has been processed at a deep enough level to allow us to suggest these items had gone beyond the "limited incidental acquisition" the authors mention. This might lead us to plan for a conceptual replication of the study with a more longitudinal design.

» Activity 20

Look at pp. 207–213 in the key paper by Bitchener & Knoch.[6] Decide what constructs have been used *and* how these have been operationalized. If you were going to carry out a conceptual replication of this study to further test the underlying theory or hypotheses, how might you operationalize some of these constructs differently?

The Limits of Conceptual Replication

As we pointed out at the beginning of this section, a conceptual replication takes us some way along the "replication continuum", further away than all those studies we have looked at. Thus, any similarities or differences in outcomes that we wish to draw with that original study must be tempered with the knowledge that we are no longer making direct comparisons with the original procedures and results, but rather with the possibilities for extending its hypotheses or building further upon the theory behind it.

Indeed, because of this, many have questioned whether all this is really "replicating" at all. First, how sure can we be that the original study and the conceptual replication of it are actually observing the same phenomenon – remember you might well have chosen to provide a different definition of the main construct used. Second, we must be wary of the conclusions we draw from conceptual replications. If two studies conclude similarly after using two very different procedures, we might suggest our work has replicated the original "conceptually" to a particular degree. However, if our study did not come to

the same conclusion, can we say we have "conceptually not replicated" our original? Obviously not, as that would not be the basis of comparison in terms of procedures and/or analysis, etc to satisfy such a conclusion. Moreover, as we will discuss later, there is an apparent publication bias toward successful outcomes in replication, and a bias will quickly be presented toward our reading only of such work, rather than "failed" replications.

Having said all this, we end this section encouraging you to undertake well-planned and well-argued conceptual replications. They can be a more high-risk undertaking than close or approximate replications in the sense that failure to replicate will leave us with little or anything to say about the original. However, anything that results in our better understanding of the extent or limits of a hypothesis or theory has to be welcomed. And there are payoffs: when we are presented with a group of conceptual replications of a particular study which tend to reinforce the original hypothesis or clearly build upon its underlying theory, we inevitably need to sit up and listen, for something is clearly emerging as a result of this work.

We have suggested in this chapter that replication in AL exists on a continuum from the more closely allied to the original study through to the more distant. Our most "distant" (conceptual) replication is, nevertheless, still seen to be of potential value as a contribution to a debate about a study. It takes us beyond the confines of the study in question and potentially presents evidence for the success of what has been tested across many more contextual variables.

The result of a well-designed, cumulative program of replication research is ideally going to result in research which has been tested or validated to the best our circumstances will allow. It will help us understand how far the observed outcomes are generalizable, and also indicate how far these effects are robust to variations in contexts and/or conceptual modification.

Notes

1 Polio, C. (2012). Replication in published applied linguistics research: A historical perspective. In G.K. Porte (Ed.), *Replication Research in Applied Linguistics* (pp. 47–91). Cambridge: Cambridge University Press.
2 Bahr, H., Caplow, T., & Chadwick, B. (1983). Middletown III: Problems of replication, longitudinal measurement, and triangulation. *Annual Review of Sociology*, 9, 243–264.
3 Rosenthal, R. (1991). Replication in behavioral research. In J.W. Neuliep (Ed.), *Replication Research in the Social Sciences* (pp. 1–30). Newbury Park, CA: Sage.
4 Textual enhancement is a form of modifying visually those parts of a printed text which include a targeted syntactic structure for the purpose of instruction. The aim is to bring the learner's attention, while s/he is focusing on the meaning of a stretch of discourse, to the targeted structures and to how they are used. It is hoped that textual enhancement will promote the learner's noticing of the form and will help them acquire or comprehend them.
5 See Introduction (pp. 10–11) for full reference.
6 See Introduction (pp. 10–11) for full reference.

6

EXECUTING AND WRITING UP YOUR REPLICATION STUDY

Research Questions and Methodology

6.1 Introduction

Executing a replication study and writing it up for journal publication is our focus in both this chapter and Chapter 7. To achieve this goal, we will pay particular attention to critical reflections on study design, methodological procedures, and routine checking of results – as discussed in Chapters 2–5. Because describing accurately what went on is such a critical part of the research process, we dedicate large portions of Chapters 6 and 7 to the writing up of replication research, including suggested writing models with examples from published replications in high quality journals.

Just as we understand that replication can differ from other types of empirical research, the writing up of a replication study also includes unique features that we need to be aware of. For example, since our replication is likely to differ from the original study, we need to make explicit *what aspect(s)* of the original study have been changed, *why* we have changed them, and *the way(s) in which* our changes have been executed. As we will see, detailing the procedures and rationale for our replication becomes necessary not only for interpreting our findings in light of those from the original study, but also in terms of contributing to the larger academic discourse (see Appelbaum *et al.*, 2018).[1] Researchers carrying out subsequent replications that build on your own will need to clearly understand how you approached replication, and why.

Understanding journal expectations for replication research is therefore important. Let us begin with surveying journal requirements for the writing up of replication research. We will look specifically at one journal in second language acquisition, and another from an unrelated social science field.

» Activity 21: JOURNAL REQUIREMENTS FOR THE WRITING UP OF REPLICATION RESEARCH

A leading journal in our field that has dedicated space to replication research is *Studies in Second Language Acquisition*.

Examine the submission guidelines for replication studies in *Studies in Second Language Acquisition*, as described on the journal's webpage www.cambridge.org/core/journals/studies-in-second-language-acquisition/information/instructions-contributors. List the requirements for replication research, and then compare these against submission requirements for a research article.

Next, compare the above guidelines for replication research with those from a journal in the field of information systems, *AIS Transactions on Replication Research* (http://aisel.aisnet.org/trr/). What major differences do you note?

In this chapter, we focus on your replication study's research questions and methodology. Chapter 7 deals with the analysis, results, discussion, and conclusion. In both chapters, we are going to execute and write up a close replication of one of the key papers, Bitchener and Knoch (2010). Two published replication studies of second language acquisition research will serve as our models.

The first of these is Eckerth's (2009)[2] replication of Foster's (1998)[3] study on the negotiation of meaning, published in *Language Teaching*, which followed the original study's methods of data collection and data analysis. The modified variable was target language (L2 German), plus addition of a new data collection instruction (stimulated recall). The second is McManus and Marsden's (2018)[4] replication of McManus and Marsden (2017)[5] on the effectiveness of L1 explicit instruction for L2 grammar learning, published in *Studies in Second Language Acquisition*, which very closely followed the original study's methods of data collection and data analysis. The modified variable was type of explicit instruction (L1 practice without L1 explicit information).

Before we begin, it will be useful to briefly review some of the earlier content from Chapters 3 and 4 about critiquing a research study. In planning our replication, we need to reflect on central aspects of the original study's research design.

» Activity 22: CRITIQUING RESEARCH DESIGN

Chapter 4 offers some guided reflection on important design features of published research. Take a look over section 4.1 (pp. 48–65). Based on your rereading, answer these questions:

- What is the impact of scant reporting of the original study's data sample?
    - How could you execute a replication if you know little about the data sample?
- In what ways might a study's findings be influenced by "statistical power"?
    - How could a replication be useful in addressing "low power" in previous research?
- In what ways can data measurement influence a study's findings?
    - How could a replication address measurement effects?
- How might extensive use of *p* values influence a study's conclusions?
    - In what ways could a replication address the limitations of interpretations based only on *p* value?

6.2 A Close Replication of Bitchener and Knoch (2010)

We will refer to Bitchener and Knoch (2010, henceforth B&K) as the "original study" because this will be our source study for replication. A large amount of our previous discussion has dealt with understanding and critiquing many aspects of this study, so many of the details should be familiar to you.

We worked closely on B&K in Chapters 3 and 5. In Chapter 3, we annotated the abstract by applying critical reading strategies (e.g., questions about research design, construct operationalization, and analysis). In Chapter 5, we drilled down into the replication process, and planned out a series of hypothetical close replications[6] of B&K. Our critical reading of B&K led us to propose four variable modifications that could be implemented in a close replication: (1) participant status, (2) amount of time spent in the host country, (3) motivation, and (4) L2 proficiency. Building on our previous engagement with B&K, and for the purposes of this close replication, our variable modification will be **L2 proficiency**.

We discussed in Chapter 5 that any replication must include a justification as to (1) why we have decided to replicate the study and (2) why we have chosen a certain variable for modification. As presented in that chapter, not only has previous research indicated that L2 proficiency level may mediate the effectiveness of corrective feedback (CF), but B&K themselves call for future research to examine the extent to which L2 proficiency differences might influence the effectiveness of CF on L2 writing.

Now is the time to refresh yourself with the justifications for replication and variable modification (Chapters 3 and 5), and then complete Activity 23.

In the remainder of this chapter, we closely review B&K's research design to ascertain the extent to which replication of this study is feasible. This analysis

will need to take into account both the completeness and transparency of the original study's methodological and procedural reporting, as well as our own expertise, and access to resources and participants. This reflection then leads to the execution and writing up of our replication study's research questions and methodology.

» Activity 23: JUSTIFYING THE REPLICATION AND VARIABLE MODIFICATION

It is not wise to replicate for the sake of replication. We need to be able to convincingly motivate our replication, as well as any changes to the original study that we implement for our replication. Your critical reading of B&K (in Chapter 3) has led to a number of questions based on the abstract, which were further fleshed out in the reading of the study's research design.

- What major questions arose after your reading of the abstract?
  - Think about: level of detail, findings reported, links between methodology and results.
- How did you address the concerns you had in the abstract?
  - Did you look elsewhere in the article? What sections in particular?
- Did a closer reading of the research study answer your original concerns?
  - What remaining concerns did you have?
- To what extent did the discussion and conclusion raise the same questions?
  - Were there unanswered questions and/or explanations?

6.3 Reviewing Research Design for Replication Feasibility

Let us begin by reminding ourselves of the task at hand. First, our replication requires us to follow as closely as possible the original design as described in B&K. Second, the variable we will intentionally modify will be L2 proficiency, which, as suggested in Chapter 5, will be "upper-intermediate". In short, our aim is to repeat as closely as possible B&K's study except for a difference in L2 proficiency. Any major modifications other than proficiency are likely to change the nature of our close replication.

First things first, we need to verify whether it is feasible for us to execute this close replication of B&K. Unless we already know, we will need to find answers to the following questions, which essentially address the basic research design

features of the original study: what is the sample, what was the context of the data collection, and what materials were used to collected data?

- Do I have access to a data collection site that runs an academic writing course for international university students?
    - Possibilities may include Intensive English Programs, Pre-Sessional courses, and English for Academic Purposes courses.
- Does that course include at least 63 ESL learners, aged 18–20 years old, who are mostly from East and South Asian countries?
- Does that course last at least ten weeks?
- Can I obtain – or satisfactorily recreate – the data collection materials used in B&K?
- Teacher participation is needed – so can I get agreements from on-site teachers?

You will want to ask yourself questions similar to these before proceeding any further because our answers may well influence what happens next.

For example, if we don't have access to a data sample similar to that described in the original study, we might need to rethink our variable modification in order to execute our close replication. Imagine, for example, we don't have access to 18–20-year olds, or maybe the students are *not* mostly from East and South Asian countries. Because we are at the planning stages of our close replication, we can still implement change. It is important to remember, however, that our variable modification has to be motivated by our critique of the study. For example, assuming we decided to modify age because we didn't have access to 18–20-year olds, what might our rationale be (see Chapter 5)?

A further consideration might be whether minor variable modifications are involved. If so, could you still proceed? For example, imagine you meet all the design features as presented above, but your participants are in an Intensive English Program in Scotland, whereas the original study was conducted in the US. In such a case, you would want to make a note of why you consider such a difference to be a minor variable modification. Then, you would want to ensure that you return to this matter in your discussion of the results because it is possible that this context difference has contributed in some way to your findings.

Before proceeding, let us take stock. We have asked important questions to determine whether it is feasible for us to execute this close replication. These questions addressed the nature of the data sample (i.e., the study participants), the context of the data collection, and the materials used to collect the data. Our answers to these questions may determine whether or not we are able to execute a close replication of the original study. For our purposes, the results of our preliminary questioning were positive.

6.4 Research Questions

In this section we examine B&K's research questions. Our aim is to determine if we should use the exact same research questions as in the original study, whether adaptations are needed, and/or whether new research questions should be added.[7] As in earlier chapters in this book, we need to critique and understand B&K's research questions in order to determine how to proceed in selecting the research questions for our close replication.

6.4.1 Understanding and Critiquing the Original Study's Research Questions

B&K's research questions are listed in a section titled "Aims" (p. 211 in the published article). That section ends with their research questions. Before we proceed, complete Activity 24.

» Activity 24: B&K'S AIMS

B&K summarize their study's aims and research questions on page 211. What are the study's main features?

Points to consider:

- motivation for the study;
- research design and tests;
- instructional treatments;
- measures of improvement.

Below are B&K's research questions:

> 1. Does advanced learner accuracy in the use of two functions of the English article system improve over a 10-week period as a result of written CF?
>
> 2. Does advanced learner accuracy in the use of two functions of the English article system vary according to the type of written CF provided?
>
> *(Bitchener & Knoch, 2010, p. 211)*

A cursory reading of research questions 1 and 2 indicates that question 1 is about whether performance improves over time, while question 2 is about the effect of the type of written CF on performance over time. As such, question 1 examines performance over time irrespective of the type of written CF received and question 2 examines the extent to which different types of written CF lead to different learning trajectories. In a way, these questions are cumulative:

Question 1: can CF lead to accuracy improvement?

Question 2: can different types of CF lead to equal amounts of improvement?

Let us examine each separately in the same critically attentive way we addressed the reading of the abstract and other sections of the paper in Chapter 3.

First, question 1 narrows the focus to improvement of "accuracy", although we do not yet know how "accuracy" is defined (but see the "Analysis" on p. 213). It is also as yet unclear from the research questions what "use" refers to, but we assume that it refers to writing because the sample population is "L2 writers", as described in "Aims" (p. 211). We therefore assume that B&K's dependent variable is *written accuracy* (but this is specified neither in "Aims" nor in the research questions). As regards "the use of two functions of the English article system", Activity 24 told us that B&K investigated "first and subsequent or anaphoric mentions", but we should note this information is missing from the research questions. The last component of research question 1 stated improvement "over a 10-week period as a result of written CF", which indicates that written accuracy improvement was examined over a 10-week period following the provision of written CF. The "Aims" section tells us that written CF was only provided once, between pre-test and post-test in a pre-test-post-test-delayed post-test design.

Second, research question 2 is structured in the same way as research question 1 up until the last part, which examined the extent to which written accuracy varied "according to the type of written CF provided". Research question 2 indicates, therefore, that B&K additionally examined the role of different types of written CF on written accuracy. Although B&K do not specify how many types of written CF were provided, the "Treatment" (p. 212) and "Procedure" sections (p. 213) describe three types of written CF.

Our analysis of B&K's research questions could be summarized as follows:

- Question 1: The effects of written CF on written accuracy at immediate post-test and delayed post-test.
    - An important omission in the "Aims" section is that one group did not receive any CF (the control group), which justifies question 1. If no control group were included in the design, there would be little rationale for question 1.
- Question 2: The effects of different types of written CF on written accuracy at immediate post-test and delayed post-test.

Given that B&K's participants were "advanced L2 writers", we could include a proficiency level reference in our summary. But, since proficiency level was not experimentally manipulated (i.e., all groups were of the same

L2 proficiency level, in contrast to written CF, which was different for each group), inclusion of proficiency level could be seen as optional. In which case, we would have to make sure proficiency level is defined and explained in our "Aims" section.

Now, we are at a point where we have critiqued B&K's research questions to understand the focus of their investigation. Given that our variable modification is L2 proficiency, and that our close replication is following every other aspect of B&K's study, we could take their exact research questions and replace "advanced learner" with "upper-intermediate learner". That said, our critique of B&K's research questions indicated we could add more information that would improve the clarity and precision of our research questions. In the next section, we will write up our research questions for publication.

» Activity 25: RESEARCH QUESTIONS IN A REPLICATION STUDY

Examine how Eckerth (2009, pp. 113–114) set out his replication study's research questions.

Consider:

- How are Eckerth's research questions connected to the original study?
- Are any changes made?
    - If yes, in what ways, and how are they justified?
- How are the replication study's research questions presented?
    - In a different section, quoted directly from the original study?

6.4.2 Writing up Your Research Questions

There are two parts to writing up the replication study's research questions. First, we need to *define* these questions. Second, we need to *frame* them with reference to the original study. We will again use Eckerth (2009) as an example.

Let us begin by defining the research questions for our close replication of B&K. To help get us started, we can use our summaries as sketched out above in **section 6.2**. Because all other aspects of our replication are the same as the original study except for L2 proficiency, we previously mentioned that we could use very similar wording as in the original study. Replacing "advanced learner" with "upper-intermediate learner" would give us this:

1. Does upper-intermediate learner accuracy in the use of two functions of the English article system improve over a 10-week period as a result of written CF?
2. Does upper-intermediate learner accuracy in the use of two functions of the English article system vary according to the type of written CF provided?

We also mentioned that we could include additional information to improve the clarity and precision of our research questions. For example, we could specify that "use" refers to L2 writing, that "two functions of the English article system" refers to first and anaphoric mentions. We could also add information about the post-test and delayed post-test to further specify that our data was collected at two specific time points. These additions could look like this:

1a. To what extent does providing written CF improve the accuracy of first and anaphoric mentions in written L2 English immediately after instruction (at post-test) and eight weeks later (at delayed post-test)?
2a. To what extent do different types of written CF (direct, indirect, direct + oral review) improve the accuracy of first and anaphoric mentions in written L2 English?

You will notice that we additionally added "to what extent" rather than using "does", which encourages us to ask questions about degrees of improvement instead of binary questions of "improvement" versus "no improvement" (see Cumming & Calin-Jageman, 2017).[8] You may also have noticed that we dropped L2 proficiency from our research questions, for the reasons discussed earlier: L2 proficiency is not an experimental manipulation in either B&K or in our close replication. Such a decision is optional, however. If your close replication did experimentally manipulate L2 proficiency level (e.g., low, intermediate, and high in the same study), it would be essential to include this information in your formulation of the research questions.

For now, we have a set of research questions that, despite some differences in structure, reflect the same information as in B&K, but with more precision. Having defined our research questions, the next task is to write them up and link them to the original study. The extract below is from Eckerth (2009). His write-up begins with a summary of the original study's research questions and aims. It appears under the subtitle "The original research study" (p. 113).

> Foster's original study was set up to see "'what the student in the classroom does" with the negotiation of meaning (Foster 1998: 5). For this purpose, lower intermediate ESL learners in an actual classroom were observed while they were working in small groups and in pairs on different language learning tasks. The study sought to investigate to what extent the learners would

> produce a) talk in general, b) comprehensible input, and c) modified output, and whether the variables 'task type' (optional vs. required information exchange) and 'participant structure' (group vs. pair work) would affect a), b), and c).
>
> *(Eckerth, 2009, p. 113)*

Immediately following Eckerth's summary of the original study, he provides the following information about the replication study, titled "The replication study":

> The replication study closely followed the research procedures adopted in the original study. As will be specified in the following sections, the relevant parameters such as participants (3.1), setting (3.2), tasks (3.3), data collection procedures (3.4) and data coding (3.5) were identical or closely comparable to those in the original research, whereas the stimulated recall methodology (3.6) has been added to the research design.
>
> *(Eckerth, 2009, p. 114)*

» Activity 26: THE ORIGINAL STUDY AND REPLICATION IN ECKERTH (2009)

Using the previous citations from Eckerth (2009) describing the original study and the replication, answer these questions:

- Why does Eckerth cite the original study's research question?
- What information does Eckerth include in summarizing the original study's aims?
- How does Eckerth indicate similarities and differences between the original study and the replication?
- What particular expressions indicate original-replication similarities?

Eckerth begins by summarizing the original study, immediately followed by key features about his replication, with important information about how the research design was changed, and what additional tests were added. You will also notice that Eckerth cites the original study's research question. Importantly, Eckerth did not add his variable modification (a different target language) to the research question, but this is presumably because there was no explicit reference to L2 English in the original study's research question.

Because the variable we modified was included in B&K's research question, it would have been misleading to cite the original study's research questions. Deciding whether or not to cite the original study's research question is going

to be a matter of what information is included within it and personal preference. However, connecting our replication to the original study is *not* a personal preference: our write-up should explicitly state research design similarities and additions between the original study and the replication. Eckerth's expressions such as "closely followed" and "were identical or closely comparable to those in the original research" provide continual in-text support which helps the reader understand similarities to the original study. At the same time, explicit reference to differences helps the reader appreciate new or changed aspects, like "has been added to the research design". As we shall see below, such references are a hallmark of the written-up replication study.

We provide below a suggested formulation for our close replication of B&K's research questions. We use the same structure as in B&K ("Aims" with research questions), but, similarly to Eckerth (2009), we report similarities and differences between the original study and the replication (see also Appelbaum *et al.* 2018, note 3).

> Aims
>
> This close replication followed very closely the procedures and research design as described in B&K (2010), except for a difference in L2 proficiency. Modification of L2 proficiency level was the only difference between the original study and this close replication. B&K's participants were classified as "advanced learners", but this replication recruited upper-intermediate level learners. Although the original study did not use an independent measure of proficiency to assess L2 proficiency level, we assessed upper-intermediate proficiency using the ETS assessment of intermediate proficiency in TOEFL writing (17–23) and pre-test performance (For more information, see Analysis). This replication is otherwise very closely comparable to the original study's research design.
>
> As in the original study, this replication examines the extent to which (a) written corrective feedback (CF) can improve the accuracy of upper-intermediate learners' written L2 English, and (b) the impact of different types of CF on written L2 English. The same two target features were examined: first and subsequent or anaphoric mentions. Written data was collected over ten weeks: pre-test in week 1, post-test in week 2, delayed post-test in week 10. CF was provided once only, between the pre-test and the post-test.
>
> We addressed very similar research questions as B&K (2010):
>
> 1. To what extent does providing written CF improve the accuracy of first and anaphoric mentions in written L2 English immediately after instruction (at post-test) and eight weeks later (at delayed post-test)?

2. To what extent do different types of written CF improve the accuracy of first and anaphoric mentions in written L2 English?

6.4.3 Methodology

A close replication requires that we follow the original study's methodology and procedures as closely as possible. As discussed in Chapter 2, sometimes authors are unable to fully describe their research design in the paper itself because of space limitations. That said, we need to ensure that our design and procedures are reported as fully as possible to facilitate both full evaluation of our findings and replication. Today, many journals publish supplementary materials alongside published papers where authors can include additional information helpful for understanding the research more fully (e.g., extra analyses, sample materials, coding protocols). Data collection materials can also be uploaded to IRIS (www.iris-database.org), as discussed previously, which is an open-access repository of data collection materials (see Marsden, Mackey, & Plonsky, 2016).[9]

6.4.3.1 Critiquing and Understanding the Methodology

In this section, we will be critiquing and understanding B&K's research methodology to be able to execute our replication study. Our focus will now move to the following components: participants, target structures, treatments, and instruments. For more information about research methodology in AL research, see Mackey and Gass (2016).[10]

PARTICIPANTS

In **section 6.3** we reviewed B&K's research design for feasibility. We specifically asked questions about the data sample and whether we would have access to a similar one. Let us begin by closely reviewing B&K's reporting of their data sample. To begin, complete Activity 27.

» Activity 27: B&K'S DATA SAMPLE

B&K describe their data sample on page 211, which includes information about the data collection context and characteristics of their participants. Answer the following questions:

- Where was B&K's data collected?
- What type of educational context, including course enrollments, is mentioned?
- What is the age range of B&K's participants?

- What other background information is provided about B&K's participants?
- To what extent is the reported detail of B&K's data sample helpful for replication?
- Is there additional information not reported that would have been helpful? If yes, what information, and why would it be helpful?

B&K describe their data sample and the research context in the section titled "Context and participants" (p. 211). The main attributes of their research context are as follows:

> The study was conducted in the English as a Second Language Department of a large university in the USA with 63 advanced L2 writers who were enrolled in a course entitled ''Introductory Composition for International students.'' Its aim was to prepare them for the academic writing requirements of the university.
>
> *(Bitchener & Knoch, 2010, p. 211)*

B&K's description informs us that data was collected in the US in an ESL program, with international students enrolled in a course that prepared them for academic writing. Additional information about both that university's English language entrance requirements and a description for that specific academic writing course would help us clarify the extent to which our research site matched the original study. For our purposes, however, we should ensure that our research context matches B&K's as per their description.

Moving on to the data sample itself, we know from the above citation that participants were "63 advanced L2 writers" (p. 211) and they were enrolled in a course that prepared them for the "academic writing requirements of the university" (p. 211). Information about age and background was also provided: "Most of the participants were from a range of East and South Asian countries and were in the 18–20 year old age bracket" (Bitchener & Knoch, 2010, p. 211). In summary, we know that B&K's participants were 63 advanced L2 writers of English, all enrolled in an academic writing course for international students at a university in the US. We also know that "most" of the participants were from a variety of "East and South Asian countries" and most were 18–20 years old. In general, we may well conclude that the data sample is only partially reported. For example, how should we interpret "most" in "most of the participants were from a range of East and South Asian counties", what types of topics are covered in the course the students are taking? Additional useful information about recruitment would have helped our design as well. For example, how were participants recruited? This type of information might also give us some information about possible attrition and exclusion rates. Were any participants

excluded? If so, on what grounds? In short, information about inclusion and exclusion criteria would be useful information to better understand the original study's data sample.

For our close replication, however, we need to make sure that we try and meet these descriptors as closely as possible (and if we are not able to, we need to state what is different).

Because proficiency level is an important difference between the original study and our replication, we will need a recruitment method that taps into this proficiency difference. Unfortunately, information about recruitment is not discussed by B&K and no further information is available. If our particular research context offers multiple academic writing courses at different levels of instruction based on English L2 ability and/or years of English use, it might be possible to recruit participants from a beginning instructional level (e.g., semester one) in the understanding that semester one students might be less proficient in L2 English than students at the more advanced levels of instruction (e.g., semester four). We should be clear, however, that this assumption might not be borne out in the data, and so we would need to verify it.

Because proficiency is our variable modification, an important consideration is how to ensure that all of our participants are classified as upper-intermediate. Recruiting from semester one is helpful, but that might not be enough. Although the original study did not have students complete an independent measure of L2 proficiency, all participants completed a pre-test prior to instruction. Our replication study will therefore use pre-test score as an important indicator of L2 proficiency level. We will also use TOEFL writing scores as a measure of L2 proficiency, especially because universities in the US tend to require TOEFL scores for international students. The ETS assessment of intermediate proficiency in TOEFL writing is for scores between 17 and 23, and so we could take scores toward the upper end of the 17–23 range to indicate upper-intermediate.

Now, let us take stock. We will recruit participants with characteristics similar to those of the original study, except for proficiency level. Participants will be mostly from East and South Asian countries and aged 18–20 years old. They will be international students enrolled in a university academic writing course in the US. Although B&K's data sample was 63, this results in 15–16 participants per group, if numbers were equally balanced across four groups, and so we should aim for a little more than this, and we should allow for attrition. For these reasons, aiming for a final sample of around 80 will be helpful. We will recruit students from semester one.

TARGET STRUCTURES

Because L2 proficiency level is our only variable modification, our aim is to follow as closely as possible every other aspect of B&K's research design in

executing our replication. The target structures will remain the same. We now look at these as described in B&K, as follows:

> this study investigated the effect of targeting two functional uses of the English article system: the referential indefinite article ''a'' for referring to something for the first time (first mention) and the referential definite article ''the'' for referring to something already mentioned (subsequent or anaphoric mentions).
>
> *(Bitchener & Knoch, 2010, pp. 211–212)*

Examples of the target structure are included in the description of the treatments, as shown: "**A** man and **a** woman were sitting opposite me. **The** man was British but I think **the** woman was Australian" (Bitchener & Knoch, 2010, pp. 211-212). For our purposes, in line with the original study, the target structures will be uses of "a" for referring to something for the first time (the first mention) and then use of "the" for any subsequent mentions of the same thing (the subsequent or anaphoric mention).

TREATMENTS

We will also provide the same treatments as the original study, which describes three experimental treatments (see Bitchener & Knoch, 2010, p. 212). The treatment description indicates four different groups: three treatment groups, each with a different treatment, and one control group that received no treatment. The treatments consisted of written CF on pre-test texts. As a result, each participant appeared to receive individualized (and possibly different amounts of) feedback according to their specific treatment condition. Written feedback was provided on the texts participants completed for the pre-test. Feedback was provided once only for each participant.

First, in group one, participants received direct written CF plus meta-linguistic explanations:

> Group one received direct written CF in the form of written meta-linguistic explanation that included a simple explanation of the two targeted functional uses of the definite and indefinite articles together with an example of their use. Direct error corrections were not provided. Each time an error was made, an asterisk above the error referred the writer to the following explanation and illustration:
>
> 1. Use ''a'' when referring to something for the first time.
> 2. Use "the" when referring to something that has already been mentioned.
> 3. Sometimes an article is required for functions other than these two uses.

Example

> **A** man and **a** woman were sitting opposite me. **The** man was British but I think **the** woman was Australian.
>
> *(Bitchener & Knoch, 2010, p. 212)*

As indicated above, participants in group one received direct written CF plus written meta-linguistic explanations. It appears that the same explanations were used throughout, as indicated in items 1, 2, and 3 in the above citation. B&K add that "No other forms and structures were corrected" (p. 212).

Group two received indirect written CF only, by means of error circling, as follows:

> Group two received indirect written CF in the form of error circling. Thus, the only feedback that was provided to this group was an identification of where an error had occurred. Errors could occur as a result of (1) using the wrong article (i.e. definite article instead of indefinite article and vice versa); (2) failing to use an article in an obligatory linguistic environment; and (3) using an article when no article was required.
>
> *(Bitchener & Knoch, 2010, p. 212)*

Finally, group three received direct written CF plus meta-linguistic explanation as in group 1, plus oral discussion of the written meta-linguistic explanation, as follows:

> Group three received direct written CF in the form of (1) the same written meta-linguistic explanation as group one and (2) an oral form-focused review of the written meta-linguistic explanation. The latter took the form of a 15 minute full class discussion of the written meta-linguistic explanation that the writers wanted to have clarified.
>
> *(Bitchener & Knoch, 2010, p. 212)*

In executing our close replication of B&K, we will provide the same treatments as described in the original study: three treatment groups and one control group. Each treatment group will receive a different type of written CF. Group one will receive direct written CF plus a written meta-linguistic explanation. Group two will received indirect written CF only (circling of errors). Group three will receive written CF plus a written meta-linguistic explanation as well as an oral form-focused review of the written meta-linguistic explanations lasting 15 minutes.

INSTRUMENTS

B&K describe one test instrument only, a picture description task.

A potential challenge to replication can be access to a previous study's data collection instruments. In our case, the use of a single instrument (albeit three versions of the instrument) makes the process a little easier than having to reconstruct three or four different test instruments.

As previously mentioned, a number of platforms exist for researchers to share their materials, which not only aids replication, but also increases the transparency of the research conducted and may shed some light on the nature of the findings. For our replication of B&K, we were not able to access the original study's data collection materials. We checked in a number of places, and we strongly encourage this approach. For example, is information appended to the published paper or housed on the journal's website in the form of supplementary materials? Are the materials stored on IRIS? Is there a project website that we can find? Last, writing to the authors might be a good final option. In our case, none of these methods resulted in access to the original study's data collection instrument. We think that this scenario should perhaps be anticipated and so it is helpful for the purposes of this book to illustrate that not all hope should be lost. In such cases, we turn to the original study's description of the data collection instrument.

In order to execute our replication, we sought to reconstruct the test instrument as described, as follows: "Each of the three pieces of writing required a description of what was happening in a picture of a social gathering (a beach, a picnic, and a family celebration). Thirty minutes was given for the writing of each description" (Bitchener & Knoch, 2010, p. 212). We know two pieces of information about B&K's test instrument. First, the instrument was "a picture of a social gathering". Second, there were three different pictures "a beach, a picnic, and a family celebration". In reconstructing the test instrument, we should ensure that our pictures match these descriptions.

When it comes to writing up, we will want to provide more detail about the test instrument (discussed below). The more information we can provide, the easier it will be to evaluate our replication. For example, if our replication leads to a different patterning of results, we will want to discuss the fact that we matched the original study's description of its test instruments, but that the test instrument was not the same. On publication of our replication, we would also want to make our materials freely available for use and replication by other researchers, and so IRIS is a helpful option in this regard.

PROCEDURE

B&K describe their procedure for data collection in four steps. We will adhere to these procedures as closely as possible. Participants and then teachers separately "were provided with information sheets about the study and were given the opportunity to ask questions before signing a participant consent form" (p. 213).

These information sessions were scheduled five days before the pre-test. The procedure is then described as follows:

> 1. On day one, the pre-test was administered.
> 2. Three days later, the texts were returned with written CF on the texts of participants in groups one, two and three. For group one (written meta-linguistic explanation), the immediate post-test was completed after they had been given several minutes to consider the explanation. For group two (indirect feedback), the immediate post-test was completed after the participants had been given several minutes to consider the feedback. For group three (written meta-linguistic explanation and oral form-focused instruction), the immediate post-test was completed after the participants had considered the written meta-linguistic explanation and taken part in the form-focused discussion.
>
>   For group four (control), the immediate post-test was completed at the beginning of the class.
> 3. The immediate post-test for all groups was returned 1 week after it had been written. Corrective feedback was not provided on this occasion. It was returned to the participants in order to satisfy the requirements of the university's ethics committee.
> 4. The delayed post-test was administered 10 weeks after the pre-test. The teachers had agreed to not provide any instruction or correction on the targeted forms during the interim period. It was administered without the participants being given any advance notice.
>
> *(Bitchener & Knoch, 2010, p. 213)*

We will now work through the above procedure so as to fully understand the different steps involved in data collection.

First, the pre-test was administered on day one. Although a pre-test was neither clearly identified in the description of the data collection instruments nor here in the procedure, we will assume that the pre-test involved a 30-minute handwritten description of an image, and that participants had no access to dictionaries, grammar books, or other writing aids. In executing our replication, we will use the 'beach' image as our pre-test (note that the original study did not specify which image was used for the pre-test).

Second, three days after the pre-test, participants in a treatment group (groups one, two, and three) received back their texts with written CF that conformed to their particular treatment, as previously described. B&K explained that participants were given "several minutes to consider the feedback" (p. 213). At this point, the immediate post-test was completed.

Third, participants received back their immediate post-test "1 week after it had been written" (p. 213), indicating that all groups (treatment and control

groups) completed a second handwritten image description after receiving back their corrected texts (but this is not clear from the procedure). For the immediate post-test, however, no written CF was provided. For our purposes, we will administer the picnic image as the immediate post-test.

Last, the delayed post-test "was administered 10 weeks after the pre-test". Additionally, B&K note that "the teachers had agreed to not provide any instruction or correction on the targeted forms during the interim period. It was administered without the participants being given any advance notice". (p. 213). Although not specified, we will administer the family celebration as our delayed post-test.

As mentioned above for the test instrument, we will want to provide additional information about our procedure (e.g., specifying the amount of time between receiving feedback and completing the post-test, and that the texts were handwritten). Even though these were not specified in the original study, they are important to document for subsequent evaluation of our findings and further replication.

» Activity 28: REPORTING THE METHODOLOGY IN A REPLICATION STUDY

Closely examine the methodological reporting in Eckerth (2009). What do you notice about the writing up of the replication study's methodology and how does this differ from the methodological reporting in an original research study?

Think about the following questions:

- How does the author connect the replication and the original study?
  - Do you notice any expressions or use of subtitles that help connect the original study and the replication?
- How does the replication report differences with the original study?
  - What methodological differences are described, and how are they justified?
- How does the replication report similarities with the original study?
  - What particular word choices help the reader understand between-study similarities?
- Does the replication study add any new tasks?
  - If so, how are these described and how are additions justified?

6.4.3.2 Writing up the Methodology for Publication

At this point, we have critiqued B&K's methodology, with a particular focus on the data sample, target structures, treatments, instruments, and procedures. We now discuss how to approach writing up these different components for publication (see also Appelbaum *et al.* 2018, note 3). As discussed above, and consistent with writing up the research questions, we will want to clearly describe any similarities and differences between the replication and the original study, including any assumptions we made and any additions we implemented. For example, we made some assumptions about what constituted the "pre-test", and it was not clear how participants were assigned to the different treatment groups. We would want to highlight that these aspects were not described in the original study followed by a description of how we addressed these points.

We will begin by stating how our data sample is different to B&K's sample. As a reminder, we recruited participants with very similar characteristics to the original study, except for proficiency level, and so we should begin with this information, following Eckerth (2009): "While the original study was conducted with ESL learners at a municipal college in Great Britain, participants in the replication study were students in an L2 German university course" (Eckerth, 2009, p. 114). It is then important to highlight all other aspects that are the same as in the original study, which in our case is that participants were mostly from East and South Asian countries, aged 18–20 years old, and were international students enrolled in a university academic writing course in the US. It is helpful to use phrasing like "as in the original study" or make explicit reference to the author of the original study "as in Foster's study" (Eckerth, 2009, p. 114). McManus and Marsden (2018, p. 5 – see Note 4) additionally included the following because of similarity in the original and replication's data samples: "These are very similar participant characteristics to the participants in McManus and Marsden (2017)".

The following is a suggested write-up of our data sample, following conventions from both Eckerth (2009) and McManus and Marsden (2018). The comparative language has been highlighted:

> Participants
>
> **While** B&K's (2010) participants were identified as advanced ESL writers, **this replication study's participants** were upper-intermediate ESL writers because L2 proficiency was our intentional variable modification. Although L2 proficiency level was not independently measured **in the original study, this replication** used TOEFL writing scores as an indicator of L2 English proficiency. All participants reported TOEFL writing scores between 17 and 23. We also compared our pre-test scores with those

reported in the original study, and pre-test scores in this replication were descriptively **lower than in the original study**.

In line with the original study, participants were 80 learners of English as a Second Language, enrolled in semester one of an Intensive English Program at a large university in the US. All participants followed a required course in English academic writing. The mean age was 19 (range 18–20) and, **as in the original study**, the participants were mostly from East and South Asian countries (China, $n = 25$; India, $n = 14$; Japan, $n = 14$; Pakistan, $n = 10$; South Korea, $n = 13$), with a small number of participants from Saudi Arabia ($n = 4$) and Europe (France, $n = 3$; Germany, $n = 2$). These are **very similar participant characteristics** to the participants in B&K (2010).

Since the other components of the methodology section are exactly the same as in the original study (i.e., target structures, treatments, instruments, and procedures), we will cover them together. We should add, however, that as with all other write-ups of replication research, we should ensure that we consistently note similarities and differences when they occur. For example, we previously noted that we were adding detail to the procedures in terms of which tasks were used at the pre-test, post-test, and delayed post-test (which were not specified in the original study). Also, since we were not able to use the original study's data collection instruments, we created our own based on descriptions in the original study. These are important points to note for subsequent comparison with the original study and for future replication.

In short, we must ensure that our write-up contains two features. First, when something changes between the original and the replication, we must clearly state what changed. Second, when something is the same as in the original study, we must clearly state that it was the same. In the following paragraphs, we present an example write-up of our replication's target structures, treatments, instruments, and procedures.

» Activity 29: COMPARATIVE LANGUAGE

Using our example write-up of "participants" as a guide, highlight the comparative language used in Eckerth's (2009, pp. 113–116) replication study.

What kinds of comparative language emerge, and what functions do they serve?

Target structures

This replication study's two target structures were exactly the same as in B&K (2010): use of "a" and "the" to index the first and subsequent mentions, as shown in the following example. "**A** man and **a** woman were sitting opposite me. **The** man was British but I think **the** woman was Australian" (Bitchener & Knoch, 2010, p. 212). The bolding (from original) on "a" and "the" is used to illustrate that the first mention of man is referred to with "a" because they are unknown in the discourse, whereas their subsequent mention uses "the" because they are now known subjects in the discourse. The use of "a" (for first mention) and "the" (for subsequent mention) were selected in the original study because previous research has repeatedly shown these structures to be difficult to acquire, even at the advanced levels of proficiency (Butler, 2002; Ferris 2002, 2006).

Treatments

Our treatment design exactly followed B&K (2010), who provided three types of CF treatment. Written CF was handwritten and provided on the pre-test text. We note that the original study did not specify whether texts were handwritten or typed.

First, a "direct CF group" received direct CF that included a brief metalinguistic explanation of "a" and "the" when used for first and subsequent mention (as described for "target features"). Other uses of "a" and "the", correct or incorrect, were ignored without correction. For each error, an asterisk was marked above the error with the same color marking used for all asterisks, which referred to the following explanation and example (no other explanations or examples were provided), exactly as in the original study:

1. Use ''a'' when referring to something for the first time.
2. Use ''the'' when referring to something that has already been mentioned.
3. Sometimes an article is required for functions other than these two uses.

Example

A man and **a** woman were sitting opposite me. **The** man was British but I think **the** woman was Australian.

(Bitchener & Knoch, 2010, p. 212)

Second, an "indirect CF" group received indirect written CF in the form of error circling. The same color circling was used for both "a" and "the".

As a result, indirect CF only identified the location of an error, and not the nature of the error. No meta-linguistic explanations or examples were provided. Three types of error were circled: (1) incorrect article use (e.g., "the" in place of "a"), (2) article omission when one was required, and (3) providing an article when one was not required.

Third, a "direct CF + oral review" group received the same written CF as the direct CF group, plus an additional "oral form-focused review of the written meta-linguistic explanation" (Bitchener & Knoch, 2010, p. 212). As in the original study, the oral review was a 15-minute full class review of the written CF explanation.

Finally, a control group received no treatment and completed only the pre-test, post-test, and delayed post-tests.

Instruments

As in the original study, three images were used to elicit written descriptions of what was happening in the image (see Appendix and IRIS for images used). Each image was of a social gathering: One image was at the beach, one image was at a picnic, and the last image was at a family celebration. We used the same contexts (beach, picnic, family celebration) as B&K (2010) to elicit written descriptions. Written descriptions were handwritten on paper, and participants had 30 minutes to write their descriptions. No dictionaries, grammar books, or other writing aids were available. Although we were not able to use the original study's test instruments,[11] we closely followed their description as reported in the original study: "Each of the three pieces of writing required a description of what was happening in a picture of a social gathering (a beach, a picnic, and a family celebration). Thirty minutes was given for the writing of each description" (Bitchener & Knoch, 2010, p. 213).

Procedure

We followed the exact same procedure as in the original study, as follows:

1. Five days before the pre-test, we met with all participants and then with all teachers to discuss the study and give opportunities for questions to be asked about the study. All participants signed a consent form.
2. We administered the pre-test on day one of week one. Each participant received one color copy of the beach image and were requested to write a description of it on the provided paper using the provided pen. Participants received 30 minutes to complete their writing.

3. Participants were then randomly assigned to one of four groups: direct CF group ($n = 20$), indirect CF group ($n = 20$), direct CF + oral review group ($n = 20$), control group ($n = 20$).
4. Three days later (day four, week one), treatment group participants received their pre-test texts with feedback according to their grouping: direct CF, indirect CF, direct CF + oral review.
5. On that same day, all groups completed the post-test 30 minutes after receiving their pre-test texts: handwritten descriptions of the picnic image. As in the pre-test, participants had 30 minutes to write their descriptions.
6. The delayed post-test was administered ten weeks after the pre-test: a written description of the family celebration image within 30 minutes.

As in the original study, CF was only provided on the pre-test text. Also, interviews with teachers confirmed no teaching of this study's target structures during the course of data collection ("a" and "the" in first and subsequent mentions). Participants were assigned to a treatment group via randomization, but we note that the original study did not describe how participants were assigned to groups.

Notes

1 Appelbaum, M., Cooper, H., Kline, R.B., Mayo-Wilson, E., Nezu, A.M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73.1, 3–25.
2 Eckerth, J. (2009). Negotiated interaction in the L2 classroom. *Language Teaching*, 42.1, 109–130.
3 Foster, P. (1998). A classroom perspective on the negotiation of meaning. *Applied Linguistics*, 19.1, 1–23.
4 McManus, K., & Marsden, E. (2018). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*, 40.2, 459–475.
5 McManus, K., & Marsden, E. (2017). L1 explicit information can improve L2 online and offline performance. *Studies in Second Language Acquisition*, 39.3, 459–492. doi:10.1017/S027226311600022X.
6 As a reminder, a close replication involves modification of only one major variable at a time in order to keep all other study design features as constant as possible, and therefore facilitate comparison between the original study and the replication (for review, see Chapter 5).
7 For a general introduction to research questions in second language research, see Mackey, A., & Gass, S.M. (2016). *Second Language Research: Methodology and Design*. Routledge: New York.
8 Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the New Statistics: Estimation, Open Science and Beyond*. Routledge: New York.
9 Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing Methodology*

and Practice: The IRIS Repository of Instruments for Research into Second Languages (pp. 1–21). New York: Routledge.

10 Mackey, A., & Gass, S.M. (2016). *Second Language Research: Methodology and Design*. New York: Routledge.

11 We checked the article's Appendix and the journal's webpage, as well as checking on IRIS. We then wrote to the authors, but they no longer had copies of the data collection materials.

7

EXECUTING AND WRITING UP YOUR REPLICATION STUDY

Analysis, Results, Discussion, and Conclusion

7.1 Introduction

In Chapter 6, we critiqued the research design of an original study (research questions and methodology), followed by recommendations for executing and writing up a replication study using published replication studies as models (see also Appelbaum *et al.* 2018). This chapter extends the execution and writing up of a replication study to the analysis, results, discussion, and conclusion. Many of our critiquing and reporting strategies are similar to those used in Chapter 6 because our aim in writing up our replication study will again be to systematically highlight similarities and differences between the original and replication studies, including justifications for any differences.

Our focus so far has been methodological and we have not thought very much about how a variable modification might require a new or different set of analyses. A related challenge is how to deal with partially described analytical procedures in the original study. It is for these reasons, therefore, that our critique of an original study's research methodology is important for replication: in order to make informed decisions about if and how to implement alterations and additions, we need to critique and understand the original study's analytical procedures. For example, as discussed in earlier chapters, we might want to address concerns about data measurement or statistical procedures.

» Activity 30: DATA MEASUREMENT BETWEEN STUDIES

An important question for replication research is the extent to which you are bound by the original study's coding and analytical choices. For example, perhaps normality and test assumptions were not reported,

the group sample sizes are not of equal variances, or particular codings and/or variable types are not compatible with the type of statistic used. These would be good grounds for considering alternatives.

- What might be the first step to evaluating whether different analytical procedures should be considered?
    - How does the original study justify its analyses? What are the assumptions of the tests carried out? Are there other, more appropriate alternatives to the original study's coding procedures?
- What might be the impact of implementing different analytical procedures for our replication study?
    - Would different codings compromise the essential comparability of replication? Does our decision have to be binary: replace/not replace?
- You might decide that a different way of coding the data is critical, and therefore decide to conduct two analyses: one as in the original study and one following recent research in the field.
    - Would you report both sets of analyses? Would you compare the different analyses? How might this approach advance our understanding of the original study's findings?

As with our general approach to replication, it is important that we aim to follow the original study's analysis as closely as possible, but to be aware that a particular coding decision and/or data measurement technique might lead to potentially different conclusions. Publishing additional analyses in a journal's online supplementary materials can be one option to address concerns about coding decisions and/or data measurements (as in McManus & Marsden, 2018, for example).

7.1.1 Analysis

A study's analytical and statistical procedures need to be detailed enough for us to understand and evaluate a study's findings. The reader should be clearly aware of how the data was coded and analysed, including any justifications, as well as the guidelines for interpreting findings. In this section, we critique B&K's analysis, as reported in the section titled "Analysis" (p. 213). The purpose of our critique is to understand how the original study's data was analysed in order to replicate these procedures as closely as possible. However, we must also be aware of the impact of particular analytical decisions on outcomes, and so we will also consider alternative procedures, if appropriate, which might increase the transparency and robustness of the analysis. When it comes to writing up our replication study's analysis, we must also ensure that we fully document any changes followed by justifications for them.

7.1.1.1 Critiquing and Understanding the Analysis

B&K state that "the two targeted functions alone were identified for each text" (p. 213), indicating that their analysis is based only on "a" and "the" when used for first and subsequent mentions in the pre-test, post-test, and delayed post-test texts. An important uncertainty in the analysis description, however, includes the identification and coding of "failing to use an article in an obligatory linguistic environment" (p. 212).

In executing our replication, we would want to highlight that the analysis of "failing to use an article" was not clearly described in the original study (assuming further information was not available). We would also want to describe how we addressed this potential data identification and coding difference.

Perhaps the most important aspect of the data analysis concerns the operationalization of accuracy, described as follows:

> Accuracy on each occasion was calculated as a percentage of correct usage for all occasions where the grammatical structure of the sentence written by the student required it (see Ellis & Barkhuizen, 2005, for a discussion of obligatory occasion analysis). For example, on any one text, three correct uses of the targeted form from ten obligatory occasions meant a 30% accuracy rate.
>
> *(Bitchener & Knoch, 2010, p. 213)*

Thus, the original study calculated accuracy in terms of suppliance in obligatory contexts (see below). In B&K's example, each text would receive a single accuracy score. We think that two aspects of this accuracy operationalization could be improved and would warrant further consideration before replication.

First, accuracy of "a" and "the" are merged into a single analysis, but the study repeatedly presents "a" and "the" as "two functions" (see research questions, p. 211). Thus, although "a" and "the" are understood to be two different functions with correspondingly different forms, the analysis appears to disregard this distinction by analyzing them together. An important consideration at this point is therefore whether to (a) follow the original study's analysis or (b) change the analysis to reflect this potential confound.

Second, operationalizing accuracy in terms of suppliance in obligatory contexts was first questioned by Pica (1983)[1] because this method uses only accurate use in the calculation of accuracy (accurate use/obligatory contexts). Target-like use, however, calculates accuracy based on both accurate and inaccurate use. In the example above, for example, accuracy was calculated at 30% because three out of ten uses were accurate (3/10*100). In contrast, a target-like use analysis would use the same information (3 accurate uses out of ten), but calculate accuracy as 17.6% (3/(7+10)*100) because *in*accurate usage is a component of the calculation. These different calculations have important consequences for

determining accuracy, and in executing a replication we would want to consider the potential impact of these different operationalizations. It would seem to us that an operationalization of accuracy that accounts for both accurate and inaccurate use might be the most insightful. As mentioned previously, conducting *both* types of analysis followed by a comparison of findings could be a helpful way to proceed. If space is limited, we could include our additional analyses in the journal's online supplementary materials (as in McManus & Marsden, 2018).

B&K presented inter-rater reliability information about the consistency of error identification (100% agreement) and error categorization (97% agreement), which demonstrated high levels of consistency among the two raters. In executing our replication, we would want to aim for very similar reliability scores in the identification and categorization of errors. We would also want to replicate this reliability scoring method: percentage agreement of two raters.

The original study next described that descriptive statistics "for each of the three tests were calculated separately for the four groups" (p. 213). Although the analysis does not specify what descriptive statistics are calculated, the results (p. 214) show that group means and SDs are used. Given that the four groups are unbalanced (three groups of 12 participants and one group of 27 participants), presenting 95% confidence intervals (CIs) would have provided important information about the reliability of the mean scores (see Chapter 4). Additionally, providing means, SDs, and CIs for the frequencies of correct and incorrect usages, as well as for "a" and "the" separately, would be even more helpful in understanding the study's findings.

Last, before moving on to the writing up of our analysis for this replication study, we will consider the statistical tests used in B&K.

A critical omission in the original study is the extent to which any **normality** or assumption testing was carried out and reported in the text. Normality of distribution is an important assumption, and any violations would require us to either (a) select tests that take account of non-normally distributed samples or (b) transform our data (see Chapter 3). Visualizing the distribution of the data is a first step to assessing normality with, for example, density plots, quantile–quantile plots, histograms, boxplots. We can statistically test for normality with the Shapiro–Wilk normality test (or the Kolmogorov–Smirnov test, depending on sample size, see Larson-Hall, 2016).[2] These procedures would give us a good amount of information about the distribution of our sample, especially important given the unbalanced sample sizes of B&K's groups (*n*s = 12, 27, 12, 12).

In order to carry out a repeated-measures ANOVA, like the original study, we need to check that our data meet the **required assumptions** of ANOVA:

(1) normal distribution of the data;
(2) equal variances;
(3) normal distribution of residuals; and
(4) equal variances of residuals (see Larson-Hall, 2016).

If our data meets these assumptions, we would be in a position to carry out a repeated-measures ANOVA like in the original study, described as follows:

> Because no statistically significant differences on the pre-test scores were found (p = .314), a two-way repeated measures ANOVA was chosen to address the research questions. One-way ANOVAs with Tukey's post hoc pair-wise comparisons were used to isolate the exact points in time where differences between the groups occurred.
>
> *(Bitchener & Knoch, 2010, p. 213)*

The original study indicated that no between-group differences were found on pre-test scores, tested using a one-way ANOVA. A repeated-measures ANOVA was then selected to analyse performance over time because pre-test between-group parity was established. Even though pre-test scores were not statistically different between the groups (p = .314), we should be cautious of this result because p values can be influenced by sample size (see Chapter 4). Given that descriptive differences do exist between the groups (e.g., mean accuracy difference of 7.5% between Groups 2 and 3), we might want to consider carrying out a repeated-measures ANCOVA, which would allow us to better account for pre-test differences (see Larson-Hall, 2016).

We would also want to **specify our dependent variable** (accuracy) and **fixed factors** (time and group). B&K's reporting indicates that a repeated-measures ANOVA was carried out, followed by a series of one-way ANOVAs. We would need to further specify more information here. First, that posthoc testing was only carried out when the result of the omnibus test was statistically significant. Second, that posthoc tests were examined using pairwise comparisons or planned contrasts, which included a Bonferroni correction (a p value adjustment to reduce the chances of obtaining false-positive results, see Larson-Hall, 2016). Last, it will be important for us to specify our alpha level for the statistical tests (e.g., p = .05) as well as provide information about effect size calculations (e.g., Cohen's d) and their interpretation, which were not included in the original study's analysis section.

» Activity 31: ANALYSIS IN THE REPLICATION STUDY

Take a look at McManus and Marsden's (2018) data analysis reporting. Our focus is on looking for indicators of analytical similarities, differences, and additions between the original study and the replication.

- How are analytical similarities reported between the original study and the replication study?
    - Do you notice any phrases and/or word choices that are useful to highlight between-study similarities?

- To what extent are between-study differences described and justified?
    - Is additional information provided elsewhere? For example, the journal's online materials?
- Do the authors add any new analyses, or present their analyses in a different way?
    - How is this information reported to the reader, if at all?

7.1.1.2 Writing up the Analysis

As we write up the analysis of our replication study, it is important to note similarities and differences between the original and replication studies, following our model from Chapter 6 (see also replication reporting guidelines in Appelbaum *et al.* 2018). Before presenting our model for writing up the replication study's analysis, we briefly discuss the different ways we can account for similarities, differences, and additions between the original study and the replication. We also discuss some of the ways that we can compare findings between studies.

» Activity 32: REPORTING ANALYTICAL PROCEDURES IN THE REPLICATION STUDY

Take a look at Eckerth's (2009) reporting of "transcription and coding" (pp. 116–117), and the ways in which similarities and differences between studies are indicated.

- What aspects of the data analysis are the same between the studies?
    - What word choices are used to convey between-study similarity?
- What aspects of the data analysis are different between the studies?
    - What word choices are used to convey between-study differences?
- How does Eckerth (2009) approach a potentially different coding scheme for measuring speech production?
    - Why did he choose to use the original study's coding in the end?

As with our previous discussions, highlighting similarities between studies can be relatively straightforward. For example, Eckerth (2009, p. 116) noted that "in order to ensure comparability, the replication study used the same coding procedures as the original investigation", while McManus and Marsden (2018, p. 465) noted that "these analyses mirrored that of McManus and Marsden (2017)". In both cases, the authors described the procedures undertaken (e.g., coding, statistical

procedures) followed by a summary statement noting that the replication closely followed the analysis as described in the original study.

Addressing differences, however, has the potential to be more difficult, depending on their nature. Because it is important for the replication to remain comparable to the original study, any major analytical changes may reduce the degree of comparison with the original study. In our critique of B&K, we noted a variety of changes that we could implement. For example, we noted that B&K's analysis merged "a" and "the" into the same analysis, and that analysing them separately would be both (a) consistent with the rest of the paper's claims that "a" and "the" are functionally distinct and (b) reduce potential confounding influences on the study's conclusions.

» Activity 33: IMPLEMENTING AN ANALYTICAL CHANGE

We have indicated at least two potential changes to B&K's analysis: (1) analysing "a" and "the" separately and (2) operationalizing accuracy in terms of target-like use.

Assume that you want to implement either or both of these changes in your replication. How could you implement changes to the analysis without majorly compromising comparability between the original study and the replication?

Points to think about:

- What justification do I have for considering an alternative analysis?
- How will I report an alternative analysis?
- Has previous research in this area implemented the analysis I am considering?
    - If yes, what was the impact of that result on the study's findings?
- If I implement two analyses, how will I interpret the results if they contrast?

Having considered an alternative analysis, we could begin by simply stating what we understand the problem to be and its potential impact on the findings. For example, by merging performance on "a" and "the", there is the potential that we are concealing different routes of development (e.g., learners are more accurate using "a" than using "the"). However, we might then ultimately not implement the change in order to remain consistent with the original, following Eckerth:

> Foster, Tonkyn & Wigglesworth (2000) have suggested the "A-unit" as better suited for the measurement of spoken language and have, rather atypically, attempted to operationalize the unit in a detailed way. However, in

> order to ensure comparability, the replication study used the same coding procedures as the original investigation. Thus, all data was counted for c-units, defined as "utterances, for example, words, phrases, and sentences, grammatical and ungrammatical, which provide referential or pragmatic meaning to NS–NNS interaction" (Foster 1998: 8, referring to Pica *et al.* 1989 and Brock 1986).
>
> *(Eckerth. 2009, pp. 116–117)*

We see that Eckerth notes there might be better ways to analyse spoken language but implementing such a change could reduce comparability with the original study. As such, the potentially different analytical procedure is presented (which can be later addressed in the discussion), but the original procedure is implemented "to ensure comparability".

Alternatively, both analytical procedures could be implemented. For example, it would be permissible to include the analysis as described in the original ("a" and "the" merged) as well as a different analysis ("a" and "the" analysed separately). McManus and Marsden's (2018) replication calculated effect sizes both similarly and differently to the original study, and then presented both sets. The additional/different calculations included "within-group ES [effect sizes] corrected for the dependence (correlation) between the means" (p. 465–466), which were presented in the journal's online supplementary materials.

In our example of analysing "a" and "the" together and then separately, we could opt to include both sets of analyses in the main body of the paper, in the online supplementary materials, or as an appendix. It would be important to compare the two sets of results, however, to verify the extent to which the different analytical procedures led to a different/similar patterning of results. For example, do the two analyses converge or do they suggest different patterns of findings? In the latter case, this could be an important discussion point later on in the replication study write-up.

Although the discussion provides an important space for comparing results between studies, we can also use standardized effect sizes (e.g., Cohen's *d*) to make between-study comparisons, as follows: "Between-group ES are provided for each of McManus and Marsden's (2017) groups using the mean and standard deviation of the relevant group from McManus and Marsden (2017) as the 'comparison/control' group" (McManus & Marsden, 2018, p. 466). This comparison method used the means and SDs from the original study to draw between-group comparisons with the replication study (see McManus & Marsden, 2018, Table 4). In our case, between-group comparisons would be helpful in determining the effectiveness of the different treatments between studies, especially since our intentional variable modification is L2 proficiency. This would allow us, for example, to directly compare the effectiveness of the same instructional treatment among advanced (original) and upper-intermediate (replication) L2 writers. Furthermore, although B&K did not calculate effect sizes (a limitation discussed

further below), they provided means and SDs, which we can use to calculate effect sizes for our replication.

Now, having summarized some of the different ways in which we can highlight similarities, differences, and additions between the original study and our replication, let us proceed to writing up our replication study's analysis. Before doing so, let us take stock of the major points (and critiques) of B&K's analysis:

- "A" and "the" were identified, but it is unclear how "omissions" were examined.
- Results for "a" and "the" were merged into the same analysis, but separate analysis might be appropriate.
- Accuracy was calculated as suppliance in obligatory contexts, but target-like use is an alternative accuracy calculation.
- Inter-rater reliability was high for identification and categorisation of errors.
- Descriptive statistics (means and SDs of group accuracy percentages) were calculated at each test point (pre-test, post-test, delayed post-test).
- One-way ANOVA used to test pre-test differences.
- Repeated-measures ANOVA used to compare change over time, but unclear whether assumptions were tested.
- Only treatment groups were included in the repeated-measures ANOVA.
- One-way ANOVA used for posthoc tests, but with uncorrected alpha level leading to potentially false-positive results (Type I error, see Chapter 3).
- Alpha level not stated in the analysis section (but included later in the results).
- CIs not included.
- Effect sizes not calculated.

Using the above summary, and the models from Eckerth (2009) and McManus and Marsden (2018), we provide below an example write-up of our replication study's analysis.

> Analysis
>
> Consistent with B&K (2010), for each text at each test point (pre-test, post-test, delayed post-test), we identified all instances of "a" and "the" used to refer to first and subsequent mentions. We coded all uses as "correct" or "incorrect", scored as 1 or 0, respectively. For correct and incorrect usage, we calculated for all groups at each test point the means, SDs, and 95% CIs.
>
> Accuracy of "a" and "the" use was calculated as suppliance in obligatory contexts (%SOC) in percent (Ellis & Barkhuizen, 2005).[3] Although %SOC can lead to inflated accuracy rates because inappropriate uses are not accounted for, in contrast to target-like use (see Pica, 1983), this replication study measured accuracy using %SOC to ensure comparability with the original study. In contrast to the original study, however, we present two

analyses of accuracy. First, following B&K (2010) we analyse "a" and "the" together. Second, because "a" and "the" are understood to be functionally distinct, we present separate analyses for "a" and "the" to examine accuracy similarities and differences over time.

As in the original study, we calculated inter-rater reliability as percentage agreement. Two raters coded all of the data separately, and their codings were compared, which revealed 100% agreement on the identification of errors (accurate vs inaccurate use) and 97% agreement on the categorization of errors (first mention error vs subsequent mention error).

As all data sets were normally distributed (according to visual checking of normality and Q–Q plots, and Shapiro–Wilks tests), we present the result of parametric tests (ANOVAs).

First, we tested for parity (one-way ANOVA) at pre-test between all groups, which indicated no differences between the groups (CIs passed though zero, indicating unreliable change, Cohen's *d* effect sizes were marginal, $d = .09$, $p > .05$). Second, a 4 x 3 repeated-measures ANOVA was conducted, with Group as the between-subjects factor (Direct CF, Indirect CF, Direct CF + Oral Review, Control) and test point as the within-subjects factor (pre-test, post-test, delayed post-test). We set the alpha level at .05. If, according to a repeated-measures ANOVA, a statistically significant effect was found, pairwise comparisons with Bonferroni correction were used for the posthoc tests.

For interpreting magnitudes of change, we present Cohen's *d* effect sizes and 95% CIs for *d* for all between- and within-subjects paired comparisons (and not only statistically significant results). Within-subject effect sizes were calculated using the mean and SD of the pre-test as a baseline (and the post-test for effect sizes at delayed post-test). CIs that did not pass through zero were considered reliable indicators of change (Field, 2013).[4] Between-group effect sizes are additionally provided for each of B&K's (2010) groups using the mean and SD of the relevant group from B&K (2010) as the comparison/control group. We primarily draw on Plonsky and Oswald's (2014)[5] Cohen's *d* field-specific benchmarks for interpreting our *d* values (within-subjects: 0.60 (small), 1.00 (medium), 1.40 (large); between-subjects: 0.40 (small), 0.70 (medium), 1.00 (large)).

7.1.2 Results

Up to this point, we have closely examined the original study's research questions, methodology, and analysis. Each time we have critiqued the original study in order to evaluate its study's design, followed by considerations for executing and then writing up our replication. In a sense, much of the hard work is now complete. In this section, our focus is on presenting our results, following the analytical procedures previously presented. As with the other components of our replication, we aim to closely follow the original study.

» Activity 34: PRESENTING THE RESULTS

B&K's results (pp. 213–214) are presented in data tables and a graph. Statistical tests are reported both as running text and in table form. Before we proceed, examine the layout and structure of B&K's results. Consider the follow questions:

- What results are presented first?
  - Why do you think descriptive statistics appear early on in the results section?
  - What types of information do the results in Table 1 provide?
- How are Table 1 and Figure 1 related?
  - Does Figure 1 present any new or different information from Table 1
- How are the ANOVA results presented and interpreted?
- What do you notice about the reporting of the posthoc tests?
  - What is the impact of summarizing the posthoc tests?

7.1.2.1 Critiquing and Understanding the Results

B&K first present descriptive statistics for each group at each of the three data points (pre-test, post-test, delayed post-test, see Table 1, p. 214): mean accuracy in percent and the SD. Table 1 also shows the number of participants in each group, which indicates between-group differences: 12 participants in three of the groups, and 27 participants in one group.

The data in Table 1 indicate high levels of accuracy at the pre-test (all groups over 83% accuracy), with some improvement over time for all treatment groups. This data additionally shows variation in mean differences scores between pre-test and delayed post-test (e.g., 12.25% improvement for *Written CF + Oral Review* group, but only 2.17% improvement for *Indirect CF* group). Indeed, the *Indirect CF* group is over twice the size of the other groups ($n = 27$). 95% CIs would have provided important information about the reliability of the mean scores and the meaningfulness of these point average results. Also, given the pre-test differences, gains scores would have provided clearer indicators of improvement.

This data is visualized in B&K's Figure 1, which plots each group's performance over the three data points, using group mean scores. This line graph shows each group's trajectory, thus confirming the general trend from B&K's Table 1 (i.e., all treatment groups improve, but some more than others, with some visible pre-test differences between the groups). As recently discussed in Larson-Hall (2017),[6] however, graphics presenting means only can conceal important information necessary for interpreting results. To address this limitation, she proposed

a series of "data accountable graphics", which visualize data about central tendency, dispersion, and outliers. For repeated-measures studies (like B&K), Larson-Hall recommends parallel coordinate plots which can provide information about the learning trajectories at both the group and individual levels, as shown in Figure 7.1.

The parallel coordinate plots in Figure 7.1 provide information about each individual's learning trajectory (e.g., do all participants improve, do all participants trend in the same direction, do participant scores cluster together or is there a large amount of dispersion?). In visualizing our data, we should consider using data accountable graphics that do not plot group means only. Although our choice of data visualization would be different from the original study, both graphics minimally contain the same information. The main difference is that our parallel coordinate plots would additionally visualize individual performance as well as group performance.

Next in B&K's results is the presentation of the statistical tests (ANOVAs). As previously mentioned, B&K included in their statistical analyses the three treatment groups only, excluding the control group (see p. 214). This inclusion method could be followed in order to ensure comparability with the original study, but it seems like a more appropriate method would be to include all groups to examine the effectiveness of the different treatments (a) in comparison with the other treatments and (b) in comparison with no treatment.

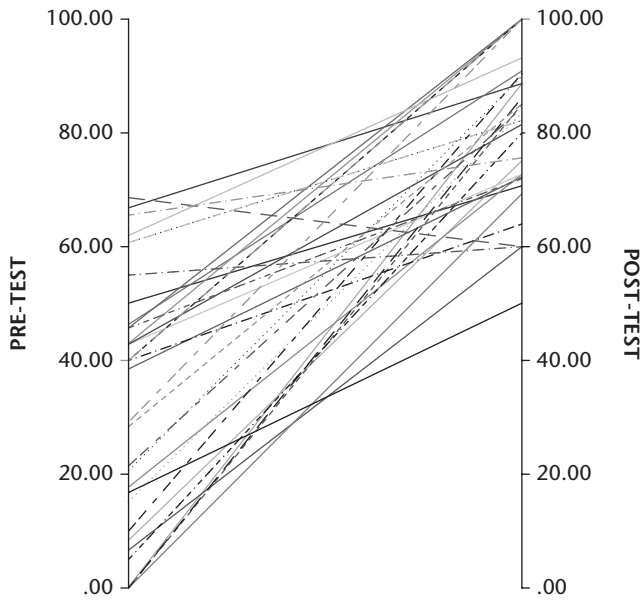


FIGURE 7.1 Parallel coordinate plot showing individual changes from pre-test to post-test.

B&K first present the results of a one-way ANOVA, used to compare pre-test scores, to examine pre-test accuracy differences between the groups. The results of the ANOVA indicated no statistically different scores between the treatment groups. That said, inspection of the descriptive statistics and the line graph suggest some between-group differences.[7] As we saw above, calculating 95% CIs and effect sizes would have been helpful in better interpreting this result. Having determined that the treatment groups' pre-test scores were not statistically significantly different, B&K's presentation continued to the results of the repeated- measures ANOVA in table 2 (p. 214).

B&K state that "there was no significant interaction between Time and written CF type", although the result of that test indicated a very strict interpretation: $p = .051$, with an alpha level of .05. It seems to us that this borderline effect should not be dismissed quite so quickly, and the use of effect sizes and inspection of CIs would indicate the extent to which the interaction between Time and CF was not a meaningful result. Furthermore, the main effects of CF type and Time were both statistically significant, which indicated "significant differences between the three groups and significant differences over the three testing times in terms of accuracy" (p. 214).

As we know, a series of one-way ANOVAs were conducted as posthocs (rather than use of pairwise contrast and/or planned comparisons, see Larson-Hall, 2016), in which an adjustment appears not have been made to the alpha level (e.g., Bonferroni correction) to account for the multiple tests being carried out on the same data set. A Bonferroni correction to the alpha, for example, can reduce the chances of obtaining false-positive results (Type I error). This case provides a strong example in favour of calculating effect sizes, but they were not included in the analysis, as follows:

> One-way ANOVAs revealed that the differences between the three groups were significant at the time of the immediate post-test ($F [3, 52] = 6.69$; $p = .001$) and the delayed post-test ($F [2, 52] = 4.67$; $p = .006$). Tukey's post hoc pair-wise comparison (with an alpha level of .05) was performed to isolate the significant differences among the three groups. These indicated that at the time of the immediate post-test, participants in the three treatment groups significantly outperformed those in the control group, but that the three treatment groups did not differ from each other. However, at the time of the delayed post-test, the participants who received indirect feedback could not sustain this improvement and therefore did not differ significantly from those in the control group.
>
> *(Bitchener & Knoch, 2010, p. 214)*

Furthermore, the results of the posthoc tests were not presented in the analysis, and so the reader is unable to examine the nature of the differences or the results of the statistical tests. We have also noted that posthoc tests appear to have

included the control group, but this group was absent from all other statistical tests. For example, B&K state all treatment groups' scores were statistically significantly higher than control group scores at post-test, but differences at the delayed post-test were not statistically significant between the control group and the *Indirect CF* group. Because the results of the posthocs were not included, we also do not know the extent to which differences between the treatment groups emerged at the delayed post-test. Effect sizes can be extremely helpful in this regard, calculated using the original study's descriptive results (means and SD).

To this end, calculating between-group effect sizes (e.g., Cohen's *d*, Hedges' *g*) is strongly encouraged when statistical tests are not reported. This process helps the potential replicator get a clear idea of the findings and their interpretation. Effect size comparisons indicated three contrasting results with the original study's claims, as follows:

1. Cohen's *d* effect size comparisons indicated pre-test differences (medium-sized effect) between the *Indirect CF* and *Direct CF + Oral Review* groups:
    - *Indirect CF* vs *Direct CF + Oral Review*: $d = .75$ [95% CIs: .03, 1.43].
2. Cohen's *d* effect size comparisons indicated no reliable differences (marginal effect sizes) between the treatment groups at delayed post-test (all CIs for *d* passed through zero), contrasting claims that direct CF is more effective than indirect CF for long-term written accuracy improvement (see p. 215):
    - *Direct CF* vs *Indirect CF*: $d = .57$ [95% CIs: -.13, 1.35];
    - *Direct CF* vs *Direct CF + Oral Review*: $d = .20$ [95% CIs: -.61, 1.00];
    - *Indirect CF* vs *Direct CF + Oral Review*: $d = .41$ [95% CIs: -1.08, .29].
3. Cohen's *d* effect size comparisons at delayed post-test indicated higher accuracy scores in the Indirect CF group than in the control group (medium-sized effect), thus contrasting B&K's claims that "the participants who received indirect feedback could not sustain this improvement and therefore did not differ significantly from those in the control group" (p. 214):
    - Indirect CF vs Control: $d = .79$ [95% CIs: .08, 1.48].

    In sum, B&K's reliance on group means and strict *p* value based interpretations have arguably presented only a partial understanding of performance. When writing up our analysis we will want to take into account the following, and aim to present our results in a way that both advances the original study's presentation and addresses limitations (many of which are presented in our Analysis section):

- Group size differences
- Presentation of group means and SDs only

- Data visualized using group averages on a line graph
- No reporting of normality, or whether assumptions met for ANOVAs
- **Control** group excluded from repeated-measures ANOVA
- ANOVA posthocs conducted without Bonferroni adjustment, which could result in false-positive conclusions (Type I error)
- Posthoc results not presented
- **Control** group included in posthocs, but excluded from omnibus tests
- Results interpreted using *p* values only

As discussed in section 7.1.3 (Discussion and conclusion), it is possible that additional descriptive/statistical information may result in different sets of conclusions. If we think that different analytical procedures have led to contrasting results, then we should draw attention to this point in the discussion.

» Activity 35: RESULTS REPORTING IN THE REPLICATION

Critically read though Eckerth's (2009) reporting of the results (pp. 118–120). Pay particular attention to reporting strategies and the structuring of the results.

Think about:

- What is the purpose of Eckerth's first paragraph in the results section (p. 118)?
- To what extent do the original study and Eckerth's replication follow the same presentation structure?
  - How do we know this?
- What is the purpose of Eckerth's use of subtitles in the results section?
- How does Eckerth summarize similarities and differences between the original study and the replication study?

7.1.2.2 Writing up the Results

Our critique of B&K's reporting of results indicated a few instances where fuller reporting would have been helpful in understanding the patterning of results, including, for example, data accountable graphics, effect sizes, CIs, reporting of posthocs tests, and less reliance on *p* values for interpretation. In writing up the results of our replication study, we can follow the same structure as the original study and choose to include additional information as necessary. However, it is important to remember that we should make sure we provide the same types of information and analyses for comparability (for additional suggestions see Appelbaum *et al.* 2018). For example, when presenting descriptive statistics, we should minimally present means and SDs. In the sections that follow, we look to our published replication studies as models (Eckerth, 2009, McManus &

Marsden, 2018). Before proceeding to the writing up of our results, let us first confirm the analyses we will present (as a reminder, see also section 7.1.1.):

- Table presentation of group means, SDs, and 95% CIs at pre-test, post-test, and delayed post-test.
- Data visualization using parallel coordinate plots.
- Include all groups in statistical analyses.
- Present results of all statistical tests (omnibus and posthocs).
- For all tests, provide *p* values, 95% CIs for the test statistic, effect sizes, and 95% CIs for effect sizes.
- Interpret our statistical tests using effect sizes, and following interpretations presented in "Analysis".
- Calculate between-group effect sizes using the mean and SD of the relevant group from B&K as the comparison/control group.

As the above list suggests, we are largely following the same results presentation structure as the original study. Indeed, Eckerth (2009) also followed the structure of the original study in presenting his findings, as follows:

> In line with the structure of Foster's paper, the data will be presented according to the research questions posed earlier: language production (Tables 2 and 3 in the e-supplement, section 4.1), comprehensible input (Tables 4 and 5 in the e-supplement, section 4.2) and modified output (Tables 6 and 7 in the e-supplement, section 4.3).
>
> *(Eckerth, 2009, p. 118)*

Following the same presentation of results structure as the original study helps between-study comparison. This approach might not always be possible, and it may well depend on the nature of the replication and the extent of variable modification. Following both our previous critique and this model from Eckerth, we could describe our reporting of results as follows:

> In line with the structure of the original study, our results will be presented in a very similar manner to B&K: descriptive statistics (Table 7.1), data visualization (Figure 7.1), statistical tests, and effect size comparisons between the original study's groups and this replication study's groups (Table 7.2).

We are thus preparing our reader for a very similar presentation order/structure. However, as we know, that will not be sufficient because we will want to indicate when the original study and replication study's results are similar and when they are different. Table 7.1 is an example of how we could present the descriptive statistics from both the original and the replication, following the same general table structure in the original study:

TABLE 7.1 Descriptive statistics for percentage group mean accuracy scores (mean, 95% CIs [LL, UL], SDs) at pre-test, post-test, and delayed post-test

Group	*N*	*Pre-test*		*Post-test*		*Delayed Post-test*	
		M [95%CIs]	*SD*	*M [95%CIs]*	*SD*	*M [95%CIs]*	*SD*
Direct CF							
B&K	**12**						
Replication	20						
Indirect CF							
B&K	**27**						
Replication	20						
Direct CF + Oral Review							
B&K	**12**						
Replication	20						
Control							
B&K	**12**						
Replication	20						

Note. Bold = results from the original study, B&K (2010)

In Table 7.1, following the original study, we have used a similar layout and have presented the same descriptive statistics (means, SDs), but have added 95% CIs. Our table also presents our results alongside those from the original study to draw attention to any differences and similarities. Depending on how your results turned out, it would be helpful to also draw readers' attention to any important similarities/differences between the original study and the replication in the text. For example, Eckerth's (2009) comparison of language production data in the original and in the replication was noted as follows: "In relation to the overall amount of language production, the extent of output modification is rather limited, as it is also in Foster's data" (Eckerth, 2009, p. 120). If our descriptive statistics showed similarities to the original study (e.g., similar accuracy proportions, high accuracy levels), we should also indicate that information using a similar strategy to Eckerth. Our purpose remains to highlight similarities and differences between our results and those of the original study.

Our visualization of data trends may indicate differences with the original study, however. This seems plausible because we are going to use parallel coordinate plots, which show individual performance. If our graphic presented a different patterning of results to the original study, we would want to draw our reader's attention to the nature of the difference, which may be because we used a different type of graphic.

Continually making readers aware of the results in the original study as compared with the replication study is a vital component of your comparative presentation.

In both Eckerth (2009) and McManus and Marsden (2018), readers were informed how the results were different, which is essential to the replication. Highlighting a difference followed by a description of the nature of the difference "Such a result is in conflict with Foster's scores, which show the same ratio for the groups, but the reverse for the dyads" (Eckerth, 2009, p. 118); "In contrast, McManus and Marsden's (2017) L21+L1 group had RTs that were significantly slower in mismatched compared to matched trials at both Post and Delayed (medium Effect Size)" (McManus & Marsden, 2018, p. 467). Not only is a difference in result important to note, but it is equally important to indicate whether the difference is perhaps due to different analyses and/or procedures. For example, we noted that the original study's statistics excluded the control group from the omnibus tests but included them in the posthoc tests. Assuming we analysed all groups together, we would want to note this difference, as it may lead to a different type of result in the omnibus tests.

Furthermore, since the original study did not present the posthoc statistics in the results, we would want to note that point. We would also want to clarify that our interpretation is based on effect size calculations (Cohen's *d* with 95% CIs), whereas the original study's interpretations were based on *p* values.

Our last point in this section refers to summarizing findings between studies. We noted that we will calculate between-group effect sizes using the mean and SD of the relevant group from B&K as the comparison/control group, as conducted in McManus and Marsden (2018), which is a helpful way to summarize differences between studies using a standardized effect size, as shown in Table 7.2

Table 7.2 summarizes the differences and similarities between the original study and the replication study because (1) it can summarize differences between groups at each test phase, and (2) it can summarize changes by adjusting for baseline differences (i.e., pre-test differences). This latter point is important because it is a standardized means of assessing improvement between studies. Although our Table 7.2 only compares differences between the same treatment groups

TABLE 7.2 Effect size comparisons (Cohen's *d* with CIs for *d*) with treatment groups from Bitchener and Knoch (2010), and effect size changes with effects adjusted for baseline differences

B&K group	Control	Direct CF	Indirect CF	Direct CF + Oral
	vs	vs	vs	vs
Replication group	Control	Direct CF	Indirect CF	Direct CF + Oral
Pre-test				
Post-test				
Delayed post-test				
Pre-post *d* change				
Pre-delayed *d* change				

in each study (i.e., Control vs Control, for illustrative purposes), you would want to extend it to compare differences between all groups (e.g., Direct CF vs Control). This approach additionally accounts for measurement differences by using Cohen's *d*, a standardized effect size (see Cumming & Calin-Jageman, 2017). Eckerth (2009) also provides a summary table of raw quantitative results (see Eckerth, 2009, table 2, p. 120), but this would only be a helpful summary if the groups were matched (i.e., no pre-test differences) and the data was measured in the same way.

Alternatively, and perhaps more traditionally, providing a narrative summary of between-study trends can indicate the important similarities and differences between studies. A narrative summary also allows for the analytical/procedural differences to be highlighted, as follows:

> McManus and Marsden (2017) found increased accuracy and speed of interpretation of the Imparfait [the target feature investigated in that study] at Delayed following a treatment of L1 EI plus L1 task-essential practice (in addition to L2 EI plus practice). We partially replicated that original study to examine the role played by L1 practice by removing the L1 EI but retaining the L1 practice (and the core L2 EI and L2 practice). We used the original study's design, procedures, and materials.
>
> Our L2+L1prac group's results patterned very similarly to McManus and Marsden's (2017) L2-only group and tended not to pattern as well with the L2+L1 group (see Tables 2 and 4).
>
> *(McManus & Marsden, 2018, p.471)*

In sum, we critiqued the presentation of results in the original study, in order to inform how we approached the write-up of our replication study's results. In line with our previous writing up sections, we showed the ways in which you can present similarities and differences between sets of results that can help your reader better appreciate the patterning of results between studies.

7.1.3 Discussion and Conclusions

The discussion and conclusions are critical components of the replication study write-up because they bring together all of the previously mentioned threads about research design similarities and differences between the original study and the replication. Up to this point, we have not engaged in any systematic interpretation of these similarities and differences, and the discussion is where we will do that. We begin with a summary critique of B&K's discussion and conclusions (pp. 215–216), before dealing with some considerations relevant to writing up the discussion and conclusions.

» Activity 36: EXAMINING THE DISCUSSION AND CONCLUSIONS

B&K's discussion and conclusions provide a summary of the findings in response to the two research questions, as well as providing some contextualization of their findings in light of previous research. Implications for language teaching are also discussed. Critically read through B&K's discussion and conclusions (pp. 215–216), paying particular attention to the following:

- How would you describe the summary of findings?
- How do B&K link their findings to the study's research questions?
- In what order are the research questions answered?
- In what ways do B&K contextualize their findings in light of previous research?
- Are any particular findings highlighted?
- Do B&K suggest areas for future research?
- What limitations of the study are addressed?

7.1.3.1 Critiquing and Understanding the Discussion

B&K's "Discussion and conclusions" section (pp. 215–216) can be seen as structurally typical of an original research paper: the research questions are restated, relevant findings are summarized, and patterns of results are interpreted and contextualized with reference to previous work in the area. The research questions are restated one at time.

B&K begin by summarizing their finding on role of CF in general: CF feedback improved accuracy immediately after the instruction, but results at delayed post-tests indicated that only direct CF was beneficial. Relevant comparisons were also drawn with the control group, a group that received no instruction. In short, providing indirect CF was, in the long-term (i.e., eight weeks after instruction), argued to be no more beneficial than providing no CF feedback. These results are then interpreted in light of previous research (paragraph 2), and then connected to pedagogy (paragraphs 3 and 4).

A very similar approach is adopted for research question two, which addresses the effectiveness of the different types of CF. One important discussion point, however, centers on the delayed post-test results because only delayed post-testing indicated differences between the treatment groups. Although not stated explicitly, adding a delayed post-test to the research design appeared to be

an important factor contrasting with previous research that has "produced essentially inconclusive results on the relative merits of direct and indirect CF" (Bitchener & Knoch, 2010, p. 216).

In terms of further research and limitations, very few are discussed, although the authors are careful to not extend generalizations much beyond the current instructional focus (use of "a" and "the" for first and subsequent mentions):

> On the evidence provided in this study, we would not want to extrapolate this possibility to other, more complex and idiosyncratic linguistic items. Until further research investigates the effectiveness of written CF in treating such items, we cannot hypothesize what the effect might be.
>
> *(Bitchener & Knoch, 2010, p. 215)*

As previously noted, the authors call for future research examining "the relative effectiveness of all types of indirect and direct feedback when given to L2 writers of different proficiency levels" (Bitchener & Knoch, 2010, p. 216). In interpreting/evaluating the findings, however, B&K offer almost no mention of methodological factors that may limit the generalizations of their findings, as we discussed in this chapter and Chapter 6, including, for example, unequal numbers of participants in each of the groups, and little information about how "advanced" is to be understood and interpreted.

In the next section, we discuss the writing up of a replication study's discussion and conclusions. While many of the attributes of writing up the discussion and conclusions of original research apply to a replication study, there are a number of important attributes that are different.

» Activity 37: DISCUSSION AND CONCLUSIONS IN A REPLICATION STUDY

Eckerth's replication (2009) discussion and conclusions (pp. 119–124) are structured differently than an original research study (e.g., B&K).

Critically examine the layout and content of Eckerth's discussion and conclusions. You will notice some clear differences in the write-up of a replication study.

Consider:

- To what extent and in what ways does the original study feature in the replication study's discussion and conclusions?
    - Where in the discussion does the original study appear?
- Is the original study discussed before the replication? In what ways?
    - Details presented, use of quotations?

- Are the findings of the original study presented? In what ways?
    - How are details presented?
    - Are quotations used?
- How are similar results presented?
    - What particular wordings do you notice?
- How are contrasting results presented?
    - Are contrasting results explained?
- To what extent are the study's findings more broadly contextualized?
- Are areas for future research discussed?

7.1.3.2 Writing up the Discussion and Conclusions

In our previous sections on writing up the replication study we have paid close attention to systematic comparisons with the original study. The rationale behind this approach is to provide guided explanations about the nature and degree of original study–replication study differences. Writing up the discussion and conclusions is no different. An important consideration discussed below, however, is how to interpret contrasting sets of findings, which has been a focus of recent discussion in both the media and academia.

» Activity 38: "WHEN GREAT MINDS THINK UNALIKE"

In May 2016, following increasing discussion of a "replication crisis", National Public Radio **(a publicly and privately funded media organization based in Washington DC, US) produced a report about replication research, what it is, and how to make sense of a number of studies that have "failed to replicate". It offers some reflection on interpreting replication research. You can listen to the podcast here:**

www.npr.org/2016/05/24/477921050/when-great-minds-think-unlike-inside-sciences-replication-crisis.

While listening, think about:

- What factors might lead to contrasting findings between the original study and the replication study?
- How can we explain contrasting results?
- To what extent should we be aware of the research context when explaining a replication study's findings?

For the discussion, re-familiarizing the reader with the original study is an important feature of the replication study. Here, you will draw out the main design features of the original study. In short, your discussion will begin with a type of "executive summary", introduced in Eckerth (2009) as follows:

> As the main purpose of a replication study is to better understand the results, outcomes, and implications of the replicated study, it will be helpful to summarize Foster's (1998: 17ff.) findings briefly. First, based on the total scores for groups and dyads in relation to task type, Foster states (i) no recognizable pattern relating task type to language production, (ii) the most consistent occurrence of meaning negotiation during pair-work on tasks with required information exchange, and (iii) the most frequent occurrence of output modification when working in a dyad setting independent of task type.
>
> *(Eckerth, 2009, p. 120)*

Eckerth summarizes the original study's main findings in "plain" English (i.e., without statistical details) in order to prepare the ground for the imminent summary of the replication study's main findings. In our case, we would want to summarize B&K's main findings: essentially that all types of Written CF were argued to improve L2 written accuracy immediately after instruction, but that between-treatment effects appeared eight weeks later at delayed post-test. More specifically, B&K argued that no differences between the indirect CF treatment and the control group at delayed post-test indicated that, compared with direct CF, indirect CF appeared less effective for improving L2 written accuracy among advanced-level learners.

Briefly restating what the replication set out to accomplish can be a helpful follow-on paragraph, in which the questions driving this replication are stated, as well as a brief rationale for these particular questions. For example, in our case, our variable modification was L2 proficiency, and a brief statement about the relevance for this particular modification would be well received. In McManus and Marsden (2018), the general research problem is restated, followed by a unified set of aims that both the original study and the replication investigated. This can be a good way of aligning both studies' aims, as follows:

> Whereas the benefits of L2 EI and L2 practice are well researched to date, the current study addressed the role of L1 practice in L2 learning. Classroom-based evidence has suggested benefits of L1 EI (González, 2008; Spada *et al.*, 2005), but has not examined L1 practice in L2 learning. Although that research had different designs to McManus and Marsden (2017) and this replication, our findings broadly align with it, extending it to show L1 EI plus practice benefited L2 offline and online performance more than L1 practice alone. In short, L1 practice without L1 EI provided few learning

> benefits, suggesting a role for L1 EI for learning the L2 features investigated here that have crosslinguistic differences for L1 English learners.
>
> *(McManus & Marsden, 2018, p. 472)*

Arguably the next component should be a synthesis of the patterns of results between the original study and the replication study, especially the extent to which these can be explained based on the original study's arguments or in other ways. In Eckerth (2009), commonalities in (a) not substantiating previous claims and (b) the original and replication studies finding the same result are introduced as follows:

> **Neither the original nor the replication study** could confirm the overriding effect of task type (required vs. optional information exchange) on the amount of language production and meaning negotiation established by former studies (e.g. Pica & Doughty 1985; Doughty & Pica 1986). Foster explains her results as a consequence of the learners' adaptation of the tasks [. . .] With regard to dyadic task completion, such task adaptation strategies **are unequivocally confirmed by the findings** of the replication study.
>
> *(Eckerth, 2009, p. 121, our emphasis)*

Important in the above extract is **highlighting commonalities** in both what was found and what was not found. However, it might be the case that different sets of results are found. It is important to acknowledge these contrasting findings as well as provide some type of explanation.

In Chapter 6 we discussed that some minor variable modifications may be included in the design of replication studies (e.g., data collection in Scotland when the original study was conducted in the US) and so it would now be relevant to reflect on these differences in discussing contrasting and similar findings. Among things we might want to discuss here:

- Did perhaps the new data collection site result in a different patterning of results?
- Would a different data collection site be a possible explanation for the different results?
- What are the implications of these differences/similarities for understanding the phenomenon under investigation?
- Could certain research design and analysis choices in the original study, which were ironed out in the replication, be a reason for the contrasting findings?

As we noted in section 7.1.1.1., certain research design and analytic choices in B&K may have influenced the original study's findings (e.g., unbalanced groups and interpretations based on *p* values only). But by balancing the numbers of participants in each group and using effect sizes, which are not influenced by sample

size unlike *p* values, we arrive at a different set of conclusions. It is critical for us to reflect on the nature of the methodological and analytic similarities and differences in discussing the replication study's findings.

After the shared *and* contrasting findings have been presented, you will notice that the replication study tends to follow the path of discussion as presented in original research, including the interpretation and contextualization of results, which helps reengage with the replication's main research focus, as well as **addressing limitations** and **further research**. For example, McManus and Marsden (2018) remind readers that only one variable was intentionally modified, which now represents a focused area for future research:

> Although we isolated the L1 practice, we did not isolate the L1 EI and so do not conclude that L1 EI alone was solely responsible for the benefits observed in McManus and Marsden (2017). The L2+L1prac's limited development suggests that the combination of L1 EI plus L1 practice led to the L21+L1 group's superior L2 performance in McManus and Marsden (2017), but future research should isolate L1 EI to test this.
>
> *(McManus & Marsden, 2018, p. 473)*

Although other variable modifications may well be necessary to better understand the nature of the original study's findings, one potential misconception of replication research might be that it could stunt further replication. But, remember that our design only modified one variable (see Chapter 5). It is possible that our variable modification revealed more questions, thus preparing the ground for further close or approximate replications.

Indeed, we strongly recommend that suggestions for further research should specifically discuss future replications, including potential variable modifications with justifications.

Equally possible is that a finding attributed to a particular treatment may be due to other factors not controlled for in the original study that a further replication is able to tease out. For example, imagine our replication contrasted with the original study's findings. And so we replicate again, but this time also with advanced proficiency learners. Over time through multiple replications, we are going to arrive at a better understanding of the phenomenon under study. Applied to the case of CF, we might better understand how a particular treatment brings about improvements, and the extent to which some other factor(s) not controlled in the design and execution of the experiment study (e.g., maybe an aspect of the course) are at play.

Notes

1 Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition*, 6.1, 69–78.

2 Larson-Hall, J. (2016). *A Guide to Doing Statistics in Second Language Research using SPSS and R*. Routledge: New York.
3 Ellis, R., & Barkhuizen, G.P. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.
4 Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. London: Sage.
5 Plonsky, L., & Oswald, F.L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64.4, 878–912.
6 Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics 1. *Modern Language Journal*, 101.1, 244–270.
7 On this point, see below for Cohen's *d* effect size comparisons showing differences between the groups, contrary to B&K's claims that the groups were not different at pre-test.

8

DISSEMINATING YOUR RESEARCH

Research is clearly of little value if it does not get out to those who should be interested in reading it. This academic community has traditionally been reached through paper publications such as books and academic journals, and is increasingly using electronic communication, blogs, conference papers, and poster sessions to become aware of new research.

It follows that those working in replication research need to have similar outlets made available to help them disseminate their work. Limiting such possibilities only serves to maintain the community oblivious to potentially crucial additional information both regarding the original studies and their replications. The importance of replication research is also being further encouraged through dedicated research grants, such as the US-based National Science Foundation which "expects that these activities will aid in verification of prior findings, disambiguate among alternative hypotheses and serve to build a community of practice that engages in thoughtful reproducibility and replicability efforts."[1]

Together with the recent calls for more replication research in AL noted in the Introduction, we have witnessed the consequent demands for further – or more effective – outlets for dissemination and/or the incorporation of such work into existing publications. The argument has even been made that there is a place for a journal specifically dedicated to – in the present case – AL replications. This, however, runs the risk of taking the important act of replication out of the main stream of AL research dissemination and potentially marginalizing it – when our intention is that it gains a much wider readership and greater funding.

8.1 Journals

For the moment – and mainly in response to the recent increased prominence given to such work – a number of high-ranking AL journals devote space or specific

strands to the publication of replication studies. This renewed interest in replication is refreshing, and such journals should constitute your first port of call for dissemination.

Journals also have the added attraction for researchers of being the principal source of future citations and, at the same time, of providing valuable exposure in prestigious outlets. Whatever we might think of journal metrics such as the widely used "Impact Factor" (see below) as a satisfactory statistic of quality, career introduction and advancement in our field remain largely dependent on having our work disseminated in such journals. Reassuringly indeed, the American Association for Applied Linguistics (AAAL) currently issues promotion and tenure guidelines which make explicit statements about how replication studies should be assessed for promotion purposes (www.aaal.org); among these recommendations we read one which states: " . . . that high quality replication studies, which are critical in many domains of scientific inquiry within applied linguistics, be valued on par with non-replication-oriented studies".

8.1.1 Selecting a Suitable Journal: General Considerations

The last decade has seen an explosion in the number of journals which purport to cover AL in all its facets. The select few journals which traditionally covered an eclectic mix of work such as *Applied Linguistics*, *Studies in Second Language Acquisition*, *TESOL Quarterly*, or *Language Learning* have been joined by many others with more specific aims and scope (including *Innovation in Language Learning and Teaching* (Multilingual Matters), *Applied Pragmatics* (John Benjamins), *International Journal of Language and Culture* (John Benjamins), and many open access journals (https://doaj.org).

In principle, therefore, you now have within your sights a considerably larger number of journals to which you can submit your work – including more specialized ones – and which would be interested in replication studies of work within their dedicated areas. However, doing a little homework beforehand to find the most appropriate vehicle for your study will reap rewards as well as save valuable time! An all-too-common mistake which editors complain about is that even rigorous, potentially high-impact work submitted to their journals is often manifestly unsuitable for that outlet and has to be rejected without further consideration.

» Activity 39

Investigate which of these journals encourage the submission of replication research, and note down and compare any specific recommendations to authors:

Language Learning

Applied Linguistics

(continued)

(continued)

Studies in Second Language Acquisition

TESOL Quarterly

Language Teaching

Modern Language Journal

Language Teaching Research

System

RELC Journal

CALICO Journal

» Activity 40

Many social science journals outside AL have encouraged submission of replication research for some time. Here are some extracts from the advice given authors. Consider

i) why you think the underlined advice has been included; and
ii) discuss whether you feel any specific guidance below might be usefully added/applied to the advice you discovered above in AL journals.

> The *European Economic Review* will in general publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the EER, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the *EER* Web site. If the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met, the editors should be notified at the time of submission. It will be at the editors' discretion whether the paper can then be reviewed. Exceptions will not be considered later in the review and publication process. Papers that do not submit this requested data will not enter the reviewing process.
>
> ****

In the case of replication studies, ***Research & Politics*** *invites authors to consider submitting a paper that is along the lines of one or more of the following replication types:*

- **Theoretical replication:** The submitted article argues that the original theoretical model is missing at least one key element. *The missing element(s) are addressed and included in the empirical analysis.*
- **Technical replication:** The submitted article identifies faults in the original research design or analysis, *thereby arguing that the original results might not hold.*
- **Concept replication:** The submitted article questions the validity of the original study. *An alternative measurement or operationalization is proposed which* yields different substantive results.

Once a replication study has been accepted to begin the review process, two reviewers will be selected and the author of the original manuscript will be notified. The author of the original study will be approached for a rejoinder, which will be reviewed (preferably a shorter paper of 2,000 words max). The General Editor will review the response and will decide if the response will be published alongside the submitted replication study.

Research & Politics allows authors to register their replication, but this is not an obligation. Registration might be relevant for cases in which existing studies are re-analyzed (ideally by collecting new data, or re-analyzing existing datasets following closely the format of the original study), particularly of research published in this journal.

****

The following is a list of guidelines/principles that apply to replications, from the perspective of the e-journal Economics. These guidelines are intended to provide systematic guidance as to how to encourage high-quality replications. This is because replications are an important public good to the economics community, and as such, they tend to be undervalued.

The journal is committed to publishing both confirming and disconfirming replications. The only criterion is that the replication be done to a high standard of professional competency.

Overview of procedure

1. . . . First, the article is submitted to the replication section of the journal. The journal sends out the replication study to the original author(s) asking for feedback within 60 days. At the end of 60 days, either with or without feedback from the original author, the journal

(continued)

(continued)

decides whether the replication study has sufficient merit to be published as a discussion paper.

Guidelines for replicating and replicated authors

. . . The original research will generally be an article in a peer-reviewed economics journal, though research from other sources (e.g., books, government publications, etc) may be considered if that research is judged to have had sufficient impact.

The replicating researcher must first attempt to exactly reproduce the original findings. If the original results cannot be exactly reproduced, then the replicating researcher is required to contact the original author to reconcile any differences. The journal will require evidence that the replicating author has attempted to contact the original author before publishing the paper as a discussion paper. Such evidence may consist of a trail of correspondence (e.g. email) which indicates that the replicating researcher has given the original author a good-faith chance to respond and/or clarify the situation . . .

The journal is willing to publish multiple replications of an original article if the studies are carried out by different replicating authors.

****

Replication Study. Authors who wish to reproduce prior research studies (***Human Factors and Ergonomics Society***) are invited to submit a proposal to submit a replication study. This is a two-stage process. Initial submissions should be a brief (1–2-page) summary of the study to be replicated and include the following information:

1. Name and author(s) of the study that has been selected for replication. The study can have been published in any reputable journal, and is not confined to *Human* Factors
2. Why the study is worthy of replication. The reasons should be one or more of the following:
    a) The study forms the basis of an important theory, model, intervention or other significant finding in the HFES literature
    b) The study is controversial in some way
    c) The study is highly cited or often viewed (please provide numbers)
3. Who will do the replication, and whether the researchers will be from a single lab or multiple labs (multiple labs encouraged)
4. Whether the original author will be included in the research group (encouraged)

5. Anticipated number of participants and power calculations/expectations (large numbers of participants encouraged)

   If the initial proposal is accepted, authors will be invited to submit a more lengthy and detailed proposal including details of the methodology and statistical analysis to be used. Further details on what will be required will be provided on acceptance of the outline proposal. If these are accepted after preliminary peer review, the accepted documents will be pre-registered in open access (the Open Science Framework) and authors will be invited to carry out their replication in the manner indicated.

8.1.2 Selecting a Suitable Journal: Specific Concerns

If you have gone to the trouble of replicating a previous study, you would understandably now be eager to get the information out to as many people as possible! Yet within that wide objective, there may be a subset of specialized readers around the world who you might want to prioritize. It's *these* readers you need to reach first.

To do so, you need first to aim your replication study at a journal with the kind of international impact and readership you want. **Journal visibility** has, therefore, to be an early, but major, factor in your decision. Aligned to this is the availability of the publication online: these days putting out your replication study on paper only can greatly limit the number of people able to access your work.

A wide, international readership will better ensure your work gets out to as many people as possible in different teaching and learning contexts. Journal web pages often publish descriptions of both the **main foci of the publication** ("About this journal" or the like) and perhaps even its current readership, both in terms of geographical spread and interests. A quick review of papers published over the last few years might also be a useful indicator to the international nature of audience and authors.

Some publishers, such as Elsevier, already provide useful online tools for finding the best journal for your needs (www.elsevier.com/authors/journal-authors#publishing-process), albeit in this case limited to those of that same publisher.

Discover **which authors have published** in that journal and whether they specialize in a similar area to that of your replication study: this might signal a journal which would be more immediately receptive to the idea of publishing a replication study which chimes with recent papers published. Similarly, look also at the **editorial board** and any **published list of reviewers** for that journal for evidence of experts in your area of interest or whether their resumés indicate similar concerns to your own.

Once you have targeted the journal, you might want to focus more closely on the **kind of papers published** (i.e., have they published/do they publish replication studies?) as well as the participants and/or methodological procedures of the papers that have been published over the last few years in that journal. For example, if you carried out your replication using quantitative statistical procedures and you notice that the journal mainly publishes papers with qualitative approaches to analysis, you might find this journal – initially at least - unwilling to consider your research for publication.

If you dig deeper into these papers, you will likely find indicators of the **kind of methodological approach favored**. The recently inaugurated *Journal of Second Language Pronunciation*, for example, encourages:

> . . . quantitative, qualitative, and mixed-methods studies . . . topics such as intelligibility and comprehensibility, accent, phonological acquisition, the use of technology (such as automatic speech recognition, text-to-speech, and CAPT), spoken language assessment, the social impact of L2 pronunciation, the ethics of pronunciation teaching, pronunciation acquisition in less commonly taught languages, speech perception and its relationship to speech production, and other topics.

Pinpointing recently published papers in your specific subject area is useful to prove your research topic is likely to be of present interest to the readers of that journal and will increase your chances of at least an initial review. Like all academic projects, AL goes through "fashions", and journals can be expected to reflect **current trends** in research interest. Check also to see if there have been any monographic or "special issues" which might give you an insight into these tendencies.

At an administrative level, you will know that each journal also has established word limits and **information on available space** for any graphics and tables. If your replication exceeds these limits you will need to assure yourself that excess (essential) data from the replication can usually be placed online in an accessible repository on the journal's website.

» Activity 41

Investigate the websites of these journals and compare (i) the main focus specified of the publication, (ii) the editorial board (specialisms), (iii) any available lists of reviewers (often listed in year-end issues), (iv) the kind of papers published (stated and as revealed in issues), (v) the kind of methodological approach favored in research (if stated), and (vi) any information on length of papers (with references), and online repositories.

Language Learning

Applied Linguistics

Studies in Second Language Acquisition

TESOL Quarterly

Language Teaching

Modern Language Journal

Language Teaching Research

System

RELC Journal

CALICO Journal

Impact indices of articles/journals you may cite or, indeed, the target study itself are also important considerations. It is true that the impact factor remains to date the yardstick used by many governmental and non-governmental institutions to assess quality in research. Care should be taken, however, in assigning value to only one index. The AAAL promotion and tenure guidelines mentioned above also remind us that:

> There are many reasons . . . for which impact factor alone cannot adequately determine the value of a given journal: the time between manuscript submission and publication in the field of applied linguistics often exceeds the two years used in the impact calculations, reducing the impact factor of applied linguistics journals. There is also a growing realization that not all citations are necessarily positive (and thus cannot be the determinant of quality), nor does the determination of the impact factor always take into consideration the practice of self-citation.

» Activity 42

Below are a number of impact indices and factors which are typically quoted on journal websites. Try to find out more information about each one, including its strength and weaknesses, and how it is calculated. Then investigate which of these is considered of greater importance or significance in your local publishing context.

(continued)

(continued)

Journal Citation Reports © Clarivate Analytics

Google Scholar Metric

Immediacy index

Eigenfactor score

Altmetric

Scimago Journal and Country Rank (SJR)

CiteScore (Elsevier)

H-Index

G-Index

8.2 Getting a Journal Interested in Your Replication Study

As you saw in Chapter 2, there are several reasons why you might choose to replicate a study. Now that you are at the stage of submitting your work for review, we might usefully remind ourselves about these reasons – and particularly those which can make a potential journal editor and review committee sit up and pay attention to your study in the first place. You might envisage this as the dissemination equivalent of "getting your foot in the door"!

We saw in that earlier chapter that not every study carried out is "significant" enough to merit or need our attention through a replication. Nevertheless, there does remain a large body of research that has influenced subsequent theory and practice to such an extent that it becomes interesting – some would say essential – for it to be revisited in some way. Replication has provided us with one such visit.

In the explanation and justification which accompanies your submission to the journal – as well as in the literature review itself – you will initially want to cite this historical reasoning to highlight why your study should be read, and published!

One important consideration might be whether to try and publish your replication in the same journal as the original study. Our take on this question is that it might be a very good place to start, but it is not necessary. One reason to consider publishing in the same place is that the journal has already demonstrated interest in that topic, and your replication will certainly be consistent with that. However, maybe that journal is not particularly interested in publishing replications, in which case you would do well to consider a different venue that publishes on a similar topic.

Let's now look in more detail at some of those original indicators used in Chapter 2 to find an ideal study to replicate and suggest ways in which you might further justify your replication submission:

> The general topic of the original paper is one that continues to generate much debate.
>
> The original paper is one that continues to generate much debate.
>
> The original paper continues to be cited in publications.

As a potential author for the journal, you face the initial challenge of convincing both the editor and the reviewing panel that your chosen study is sufficiently important to have been replicated in the first place. You will need to be very clear about the perceived need, therefore.

A study that is not only still cited after some time in the literature, but is also a continued subject of debate is demonstrating its perceived ongoing relevance. We can assume such a study has some sustained significance for the field at least and, specifically, for that journal's readership. Your accompanying argument for publication will doubtless want to emphasize how your replication endeavors to further illuminate that significance and contribute to that debate.

At the same time, and given this continued interest in the paper, the editor could reasonably expect the study already to be "familiar" to the readers. However, your "new" approach to it will be expected to throw new light on the outcomes. You will need – in your defense of this quality – to cite this recent literature and indicate what aspects of that original study are the motive of discussion and how this has led you to decide on replicating it, and its perceived contribution.

> The original paper's findings are not consistent with previous or subsequent work in the area.

Once again, much depends here on the study in question still being cited or, at least, remaining on the radar of those working in this field. If an (historically) key piece of work stands out because its findings did *not* fit in with the general trend of those in similar studies, you might reasonably argue for the need to revisit it, either to shed light on what might have brought about the atypical outcome or simply to provide more evidence.

Similarly, we might want to note here whether other replications of the same study have yielded results which in some way also point to the need for our own "take" on the replication – particularly if these have presented inconsistent results.

> The original study is cited as one of the most significant examples in practice of a particular theory.

Theory feeds into practice and practice can feed back into theory. Both theory and practice, therefore, must be interdependent if we are to advance our

knowledge and apply it appropriately. While many a language practitioner may just cite common sense as the reason why they do something in particular ways in the classroom, these often take their lead from implicit theories, even hunches, about what has worked for them before. Or, perhaps it reflects at a deeper level some strongly held conviction about the way languages are best learned. Either way, all of us who somehow contribute to the teaching and learning of languages have much to learn from theory that is well grounded in practice.

It follows that (continuing) evidence from practice can only help refine that theory. Similarly, theory-building and refinement depend on the sound nature of the research that aims to enhance them. Better (i.e., more robust) experiments should lead to better theory. Journals should be interested in reading about a replication study that advances our knowledge about a theory which remains relevant or, for whatever reason, of interest to the field.

The original study identified limitations.

We identified earlier certain aspects specific to the original study which argue for the need to replicate. These might now usefully be highlighted in this accompanying justification to the journal editor/reviewers. Limitations noted in the original study often reveal where the original authors feel future research efforts relating to the current study need to be directed; replications can contribute to the knowledge base by tweaking those limitations in the original study and uncovering remaining pieces of the puzzle.

Our baseline here again is that you have chosen a study which remains of interest to the readership. The additional contribution to emphasize is that your replication has picked up on an already-observed constraint on the outcomes in the original and sought to correct this. In this way, the replication is seen to form part of a continuing cycle of necessary input. For example, perhaps the original researcher was unable to apply random selection in the learning context chosen and wonders whether this might have affected the results. Your taking up of this aspect directly identified by the original author as of interest immediately highlights the importance of the contribution.

The statistical analysis used on the original data can now be improved upon.

Effect size data are not presented or are not convincing.

In your accompanying document to the journal you might also want to justify the importance of your contribution by claiming, for example, how your analyses have provided greater insight into the veracity – or strength – of the claimed effect. Improving the statistical power of a study of ongoing interest which has

had no further support beyond that original can also be argued to be more urgent than another which has already seen such replications. As we have seen throughout this book, you would have tried to ensure sufficient statistical power was assigned to your replication to enable you to back up the claims you now make to justify the importance of your contribution. Similarly, you might want to argue that your replication builds on the original – adding more power – since your sample size has been significantly increased (some statisticians recommend 2.5 times the original sample size).

A good case can also be made for the contribution your replication presents by arguing that new light is thrown on previous results through the method of analysis used. While your replication will want to reflect methods and procedures as close as possible to the original study, the interval between the original publication and your own replication may also have seen the appearance of other, more sophisticated, statistical procedures which could potentially further inform this data.

Clearly, the result of a replication may just as easily *not* support the findings of the original study, and your task would now be to convince those reading it that these are just as worthy of dissemination. Although confirmatory replications bring useful extra evidence which should be of interest to the readership, an outcome which does *not* fit in with previous results may be particularly deserving of further attention. Indeed, much of the recent publicity we read in the Introduction surrounding the lack of replication studies in the social sciences in general was down to the fact that both few replications were being carried out *and* because those that failed to produce the same results as in the original were not deemed of such interest to journals.

In this scenario, you might find yourself needing to justify the "failure" in other (positive) terms. There are several possible reasons for such an outcome and these need not reflect negatively on the original study. For example, your non-confirmatory outcomes might just indicate a simple case of regression to the mean: results tend to even out over time and if a variable measured presents an extreme value the first time it is noted, it will likely tend closer to the average the next time around. Then again, you might also have received interesting insights from the original authors (see below) which suggest the existence of false negatives amongst the results – perhaps failing to find evidence of effects that with hindsight they feel might have been real and which have shown themselves to be so in your replication.

A strong argument to be made from all this is that a failure to verify an original study – or some aspect of it – is just as important a contribution to science. What we are attempting to do is participate in what we referred to earlier as the self-correcting route of science. As more and more replications of a study are presented, it figures that our knowledge (and perhaps our outcomes) change.

» Activity 43

Imagine you have a replication research study to publish with the following titles.

Search out suitable journals which you think may be best targeted for the initial submission, and then target ONE according to the criteria you read about in "Selecting a Suitable Journal" above. Finally, justify your choice by summarizing why you think the journal in question is the most suitable in each case.

1. "Grammaticality judgment tests: who knows best – native or non-native speakers?"
2. "How far does form-focused instruction benefit older second language learners?"
3. "Going mobile: evidence for the benefits of language learning apps for school pupils in China."
4. "How corrected feedback in L2 writing demotivates: a view from three bilingual schools in Spain."
5. "Task-based language learning: experiences using visual and aural media."
6. "Listen with mother: story-telling as a phonological correction tool."
7. "The impact of lexical priming on L2 vocabulary acquisition."
8. "When words fail: L2 attrition in older long-term US immigrants."

Finally, a word of warning. We have been emphasizing throughout this chapter the need for you to seek out an ideal vehicle for your work – in this case a suitable journal. What you should not do – for a replication study or any other – is "blanket-mail" a number of publications, hoping that one might take up the submission. Apart from the highly questionable procedure of submitting to many journals at once, editors are regularly in contact with each other these days and often use the same referees for papers. Submission to more than one journal at a time can easily lead to immediate rejection.

8.3 Collaboration in Replication Studies

Replicating someone else's previous work can be a delicate endeavor and needs to be handled sensitively throughout the process. Editors may initially be wary of courting controversy by publishing a paper which implicitly questions what has gone on in a previous study – and thereby the original author(s).

You can alleviate these anxieties – and further strengthen your argument to proceed – by showing how you have expressly reached out to the original

author(s) throughout the process. As we saw in a number of chapters above, establishing contact with the author at the start of the study may also be essential if you need to get as many details of the original procedures as possible (both those published and perhaps those not made available because of lack of space in the journal).

Such an initial approach indicates your willingness to receive and act upon any suggestions or responses you receive. These collaborative exchanges can only enhance the comparative worth of the final product now being submitted. Indeed, as we have just seen, more and more journals are now requiring such contact be made as part of the reviewing process.

As journals would want to avoid author–author confrontations, you can turn this groundwork into a virtue! Now that you are seeking to publish your study, any evidence you can provide that demonstrates that you have actively involved the original author/s both before starting and perhaps after completing the study can only enhance your reputation and underline your preparedness for the work now presented.

Given the recent and very public spats involving original and replicating authors (See, for example, http://uk.businessinsider.com/susan-fiske-methodological-terrorism-2016-9?r=US&IR=T and https:// www.psychologicalscience.org/observer/a-call-to-change-sciences-culture-of-shaming), it might be as well to understand the potential conflict scenario that editors face so that you are better able to defend how you have sought to assuage this in your covering note with your submission. Kahneman (2014, p. 310)[2] outlines a few considerations. First, it might be assumed that the relationship between replicated and original author is initially somewhat adversarial in that the latter may be seen in an inevitably defensive position when compared to that of the protagonist role inevitably assigned perforce to the replicating author. Second, the more recent nature of the replication in comparison to the original may also be seen to favor the "latest news" aspect of the replication over the older original work. More insidiously perhaps, a proposed paper which indicates the original outcomes may be in doubt (or even in need of fine-tuning) may be considered as potentially damaging for the original author's academic reputation. Conversely, of course, this might as easily result in editorial/referee doubt over the methodology of the replication itself!

It follows that any evidence of collaborative exchange with the original author will assuage, if not remove, these concerns. There is little reason why the replicating author should not inform the original author about what he or she is intending to do, and why. It makes sense given the scenario mentioned in Chapter 2 where limited space is available in the printed pages of journals inevitably means that more precise details about the methodology used by the original author will need to be obtained. It would be useful for a summary of that input and response to be documented or summarized in an appropriate introductory section in the paper itself. As part of its publication process, the CUP journal

Language Teaching invites the original study's author to provide a short response to the replication at the end of the published replication. Such post-facto observations are an essential part of the continuous nature of the scientific endeavor of course: if the original authors notice a difference in a failed replication that they believe might have been significant, the onus is on them to mention it and perhaps call for further replications.

Such controversy, and the importance of involving the original author in your work, should not be understood as our inviting you to hold back from undertaking replications! While that author has the right to be informed about something which derives directly from his or her own work, they are not the "guardians" of the outcomes or object of interest here, and they too should embrace the attempt to advance our knowledge of this area through what you are doing. As we suggested in several chapters, AL research has been guilty of accumulating rather than constructing knowledge in many areas, and retaining "key" findings or basing theory upon single unreplicated studies can eventually be damaging if this feeds through to the teaching and textbook end of our business.

8.4 Replication Research Ethics and "Replication Bullying"

Thus, the relationship between those replicating the original work and that work's author(s) is a potentially very sensitive one. Indeed, part of the growing interest in replication research has arguably come about precisely through the increasing number of public rows between both parties.

It is a good thing, therefore, that you are apprised of the delicate nature of the work you are undertaking and of how you should best avoid – and if needs be, handle – potential conflict.

Many of the problems seem to emanate from a misunderstanding of what replication research is all about. As we described on the first page of this book, in the Introduction, replication research comes down to asking questions of what we already know, of being skeptical of what we read and, in short, adopting the inquisitive characteristic of the scientist. Such a critical attitude to what we are told or what we read inevitably leads us to doubt these claims. It does not follow, however, that our aim in such research therefore becomes part of some kind of inquisition " . . . to debunk dubious claims or sniff out potential falsehood". The objective is rather to return to a study that interests us, "repeating it in a particular way to establish its stability in nature and eliminate the possible influence of artifacts or chance findings" (Introduction, p. 2, Porte, 2012, p. 4, note 2).

Involving the original author is a big step toward smoothing the path of the endeavor, and it should be seen that way by those who are to judge your eventual work. It is a wise thing to do – but not an obligation, of course. In Kahneman's paper, the main thrust of the commentary was that researchers should follow established etiquette in dealing with the work of other people. We would equally want to be respectful and be respected when we refer to any work in our

literature reviews or when others quote our own work. The rather singular situation in replication research comes about because we are often obliged or encouraged to contact the original author to obtain more details about the study's methodology or outcomes than are immediately available in the publication itself – often due merely to the lack of space available to journals. At a practical level, therefore, it makes sense to involve the original author as soon as one decides to set up the replication.

Conflict may arise when the outcomes of a replication fail to confirm the original findings in some way or are used to question – implicitly or explicitly – the ability or integrity of either party in the replication. Reputations matter in any profession, and both those beginning their careers and established "names" can be negatively affected by outcomes – and by the way these are expressed in the paper. Much of this questioning arises when a replication fails to confirm previous results and a spotlight then seems to shift on to that original author. It is worth remembering that journals – when they *do* publish replications – traditionally prefer those that reveal a new effect or disconfirm rather than confirm the original findings (Neuliep and Crandell, 1990).[3] To a large extent, then, journal policy might well be exacerbating the situation here (although see *Language Learning*'s previously discussed implementation of Registered Reports, https://onlinelibrary.wiley.com/page/journal/14679922/homepage/registered_reports.htm.

Where researchers are simply not used to having their work revisited and, in the best sense of the word, "questioned", there will be a natural tendency to react. And if a reaction of some kind is guaranteed, the relationship replicator–author is potentially tense from the off, as the latter feels he or she must be on the defensive. At this point the original researcher can easily feel hurt or even bullied if the resulting replication fails to confirm their original findings.

8.5 Working in a Replication Team

This leads us on to the virtues of replication research which is seen to be collaborative, both in terms of involving the original authors as well as involving **larger research** teams perhaps replicating the same research in different contexts. Apart from these obvious advantages of team efforts, journals may well find it more appealing to publish work submitted and planned as part of a wider survey of the study outcomes, particularly if this culminates in increased evidence for or against that which is being studied.

In such "multi-site" replication projects, larger groups of researchers are recruited to carry out the same replication study (or studies) at the same time and at different research sites. One of the obvious advantages – apart from the increase of available data as long as enough labs are involved – is that effect sizes can be calculated more exactly allowing for between-site variability in that effect size and more convincing evidence presented for any generalizing inferences.

One of the first attempts to do this came from the "Many-Labs" project (www.manylabs.org). The initial phase of the project involved over 30 sites and 12 countries and over 6,000 participants in attempts to replicate key findings in 13 studies in psychology.

A similar, albeit more delimited, project in AL would be a welcome addition to the literature. Setting up larger research groups who are to work on close replications of one study, for example, will require careful planning and organization, not least in terms of initial recruitment of sites and members.

» Activity 44

Imagine you are setting up replication research groups to investigate the close replications you imagined in Chapter 5: pp. 73–77). Below are some general considerations and brief notes about setting up such a research group. Think about each category and expand on the notes offered to provide yourself with useful guidance on managing the group.

- Key decisions:

  How will you decide on a target study which will interest different groups in different sites?

  How will you decide which sites to approach for collaboration? National/international?

  What will be your target profile for the participating researchers? Some people are good at getting things done, while others are natural communicators and networkers – maybe a mixture of both is helpful?

  Will each research group be independent or answer to an overarching group/management?

  What complementary skill sets might be needed? Statistical skills? Data-analysis skills?

- Attracting members: how and where to publicize the research group for maximum interest in the replication objective? (email shots – where?; online presence – where?; in person – where?; use your supervisor – how?)
- Funding: Department? Faculty? University? National? Professional AL associations?
- How will feedback/reporting take place? What should be written up? Skype?
- Scheduling work: reasonable and timely cut-off points for . . . data-gathering, analysis, and reporting back.

8.6 Presenting Your Work at Conferences

For many of those reading this book, the first outlet for your research may well be the typical 20-minute conference presentation or a poster session. You will find much advice on both such presentations on the web, but given the rather peculiar (comparative) nature of a replication study, it would be as well to summarize here the main points that need to be remembered in writing up and delivering these descriptions of your work.

8.6.1 Preparing the Text

- Although the conference paper is the first communication of the replication to the field, it is not a fully-fledged paper – but often forms the basis of such a submission. The time-limited format of the conference paper means that you are equally limited in the amount of words, and therefore information, which can be conveyed. Estimates of words vary of course, but most would agree that 20 minutes would comfortably get you around 2,000 words only. As the initial communication of any research to the field it is not ideal, as you'll want your audience to hear as many details as possible. To make matters worse, in a replication it is precisely these comparative details which are likely to be the main focus of interest in what you have done! On the plus side, you might well be presenting a close or approximate replication; the fact that time limits typically mean you only highlight one or two points may actually be an advantage. It is usually a wise move at conferences to focus on one or two aspects of your research rather than try to include everything you have achieved.
- One of your first tasks will therefore need to be of *summary* in terms of theoretical, literature background and methodology and one of *selection* in terms of results and conclusions – again highlighting those that underscore the comparative nature of your outcomes against those of the target study.

» Activity 45

Search in the *Language Teaching* (Cambridge University Press) website for the following replication study:

Johnson, M, & Nicodemus, C. (2016). Testing a threshold: An approximate replication of Johnson, Mercado & Acevedo 2012. 49(2), 251-274.

i) Read the whole paper through to get an idea of the study itself.
ii) Now read section 5.1 and select/highlight what you initially think should be included in a conference presentation of this study's methodology. Justify your choice.

(continued)

(continued)

iii) Finally read sections 5.2 through 8 and *extract/highlight* what you initially think should be included to be *summarized* in a conference presentation of this study's results, discussion, and conclusion. Justify your choice.

8.6.2 Writing the Text

Presenting a conference paper means adopting a very different style and form of delivery from that of a written publication. A rookie mistake, therefore, would be to assume that you can take the sections you extracted in the activity above and simply read them out. Rather you must prepare a different text, and one honed to the specific conference audience.

The bottom line here is to present a less structurally dense text with ideas and content in short(er) sentences and with the kind of textual (and visual if you are using visual aids) signposting which helps your audience follow your reasoning and find their way easily through your talk. Try to see this paper as only the "shop window" of your work: you will want to have your audience interested enough at the end to ask pertinent questions and then leave the room wanting to read the unabridged version!

A primary concern will be whether the replication study paper is to be presented under the umbrella of a specific conference concern, such as "L2 writing", "Bilingualism in tertiary level education", "Using corpora in SLA", or a more all-embracing conference purview such as "Conference on applied linguistics in (country)". In the former scenario, you'd want to be very familiar with that field's current primary concerns and questions and then set the scene (and conclude) your contribution by showing the link between your research and the conference's broader concerns. In the case of a presentation at a conference with a much wider and less specific purview you might need to refresh your memory through some pre-reading of recent issues of the principal journals/books and, perhaps, recent work by those invited as plenary speakers to the conference. Try to find out the current debates on theory and methods and attempt to relate these to your work. The objective here is to somehow contextualize what may appear to be a narrow concern in your replication research question within the wider issues of the field and your conference audience.

There will be similar concerns in the text itself, of course. It can be a healthy exercise, we suggest, to attempt to explain your research to someone who is completely unfamiliar with your area of work. Ask the person to note down anything that was not immediately clear to him or her. Then record what was found difficult to understand or follow, or what assumptions you made but which were not confirmed in your audience's mind.

There will likely be a need for greater frequency of textual clues to help the audience follow the argument. A reader can skip ahead or pause to read a printed section again; an oral presentation does not allow for this. So take your audience through the paper. Be aware of the importance of linking words and phrases such as "nevertheless", "although", "therefore", "it then follows . . ." as well as "text-posting" with listing words such as "first", "second", "then", and "finally" after initially establishing the structure of the paper. Insert frequent "reminders" through the presentation such as "the second research question led on from the first in that . . .", or effective ways of moving into a new section or idea such as "following on from this we now move on to . . .", or "I now move on after presenting these figures to discuss the importance of those pre-test results once these post-tests were computed . . .".

Similarly, edit for – or rid the paper of – jargon or unnecessary verbiage such as adverbs, empty adjectives, or unnecessary preamble. Consider changing from the typically (printed) passive (and detached) voice to the active (and personal). Finally, remember the central place and unique importance of comparison (and comparison language therefore!) in your replication presentation and in any visual material offered (see Chapter 7). We would expect to be reminded at regular intervals (but particularly in the discussion and concluding sections) of how your work grew alongside – and takes its significance from – the original study.

» Activity 46

Look again at the replication study from the previous activity, focusing on section *6*. Imagine you want to deliver this discussion section to a conference audience who is already somewhat familiar with the literature on L2 writing but *not* specifically "pre-task planning". Use your selected extracts from the previous activity (i.e., what you thought should be included in a conference presentation of this study's outcomes), together with some of the advice given in the previous section "Writing the Text", to write out part of your conference paper. Some initial examples are given.

"By examining the effect of pre-task planning sub-processes on the written language production of L1 writers of English, who had presumably achieved the hypothesized threshold of proficiency in English, this study used measures of fluency, grammatical complexity, and lexical complexity identical to those in the original study . . ."

"Unlike the study, which found no impact of pre-task planning on grammatical and lexical complexity and a minimal impact of organization

(continued)

(continued)

pre-task planning on L2 writing fluency, this study found no impact of pre-task planning on the participants' written language production, suggesting no support for the hypothesized threshold of target language proficiency."

We examined how the sub-processes in pre-task planning affected the output of our L1 English writers. We measured fluency, grammatical and lexical complexity in the same way as in the original study.

We saw no impact of pre-task planning on output. These results were not the same as in the original study as Johnson DID describe a small impact of organization pre-task planning on the L2 output.

8.6.3 Presenting Your Work on Posters

Once again, you will find considerable help online to prepare your poster session. Here we will concentrate on suggestions for enhancing the particular nature and impact of your replication study on those stopping by.

Poster sessions are a useful way to bring your research to the attention of large groups of people. Attendees at conferences will typically browse posters at a dedicated poster session, in the pauses between conference papers, and/or during longer breaks. Your aim, however, is to have them stop and pay attention to yours, of course! Interest in replication research is growing, and you are likely to find people stopping by having been attracted by the "novel" methodological approach, especially if you make clear that your poster is a replication (for example, including "replication" in the title).

Your poster needs to serve at least two purposes:

1. For times when you are not by your poster and therefore unable to answer questions, the poster should **provide detail about the justification and nature of your replication** to permit basic understanding of what you have done.
2. It should **provide a stimulus for discussion** when you are around to engage conversation. An important consideration is that the poster should not be crammed with detailed text. Remember: a poster is neither a journal article nor a conference paper. Balancing text, visuals, and space is perhaps the most challenging aspect of creating a poster.

Many university libraries offer presentation guidelines, including templates and presentation tips. Take a look at your university library's resources to see what help is offered there.

» Activity 47: POSTER TIPS

Penn State University Libraries offers a number of resources for planning out and executing your poster presentation. You can access those resources here:

http://guides.libraries.psu.edu/c.php?g=435651&p=2970256

Have a look through the following sections:

- General layout:
  - What do you notice about recommendations for visualization, amount of text, use of color?
- Poster creation and presentation:
  - What tools and software packages are recommended for the creation of a poster presentation?
  - What presentation guidelines are suggested for interacting at a conference?
- Sample posters:
  - In the examples provided, what do you notice about layout, use of white space?
  - What sections/components are included? How detailed is the presentation?

Now, let us design a poster for our replication of B&K (2010), guided by our critique of the original study in Chapters 5 and 6.

» Activity 48: PLANNING OUT YOUR POSTER'S CONTENT

You previously looked over some resources and guidelines for creating a poster, as well some example posters. Similarly to journal articles and conference papers, posters tend to be separated into different sections, but with an important difference: a poster aims to visualize information, with much less text than both journal articles and conference papers.

For our purposes (and as a guide), we are going to plan for our poster to include five sections: Introduction/Background, Research Questions,

(continued)

(continued)

Methodology, Results, Discussion and Conclusions. There are others (e.g., References, Acknowledgements) that we will deal with as we go.

Because all of our poster sections were examined in Chapters 6 and 7, it is going to be helpful to briefly review these and make notes of their important messages. At this stage we are just sketching out the important points of our replication study.

Think about, and make notes on:

- Background:
  - Justification for the replication, including variable modifications.
- Research questions:
  - Original study's and replication study's aims and research questions.
- Methodology:
  - Brief descriptions with examples of our methodology, as well as clear indications of any differences between original and replication.
- Results:
  - Tables and visualizations of findings, and comparisons with the original study.
- Discussion and conclusions:
  - Summary of main findings, including similarities and differences with original study, further research, and limitations.

» Activity 49: SELECTING A POSTER TEMPLATE

Now that you have some idea of the poster's content (at least sketched out), it will be helpful to choose a template. The template will help you select how much text we can reasonably include. Before selecting our template, however, we need to know what size mounting board we can expect the conference to provide. Our poster will be displayed on this mounting board. Not all conferences will use the same size mounting boards, so be aware!

For our purposes, let's assume we're going to present our poster at the AAAL annual conference (www.aaal.org).

Those guidelines might typically specify that the mounting boards for posters will be "four feet by eight feet in size", and so we must ensure that our poster does not exceed these dimensions. To fit comfortably in that

space, we will aim to select a template with the dimensions 36" tall by 48" wide (A0 size), making the orientation landscape.

Now, let's select a template. The following website provides some examples to get you started, but check your university library as well:

www.posterpresentations.com/html/free_poster_templates.html.

You'll notice that some posters have different size specifications, so just be sure to select a template that is going to fit comfortably on the mounting board. Once you have selected your template, we are ready to begin.

If you have selected a poster template from the links above (which you don't have to, but it might be easier if you do), you will also likely be provided with additional guidelines on the left and right margins (e.g., tips for making sure images display correctly). These margins do not print, so you don't need to worry about deleting them. Figure 8.1 is an example poster of our replication, filled out with content (an enlarged version can be viewed at www.routledge.com/9781138657359). We are going to work through how to get something like Figure 8.1.

Before we start filling our poster template with content, we will follow some guidelines from the previously-mentioned Penn State University Libraries link:

> Aim for a total of 300–500 words on your poster. You won't simply be pasting large blocks of text from your paper or your abstract onto your poster; you need to boil it down to the essence, with explanation and visuals as needed. Use a font size slightly smaller than your name for the section headings on your poster. The rest of the text should be approximately 28-point to 36-point font.
>
> *(Penn State University Libraries, http://guides.libraries.psu.edu/c.php?g=435651&p=2970252)*

Title

Let's start at the top. The title is arguably *the* most important part of your poster. A catchy title, with a clear and uncluttered poster layout, is going to attract people to your poster, and that is what we want! In journal articles, and to some extent in conference presentations, titles can be long and descriptive, but our poster title needs to be short and snappy to grab the attention of people passing by. We want to avoid a title that is anything greater than one line.

In Figure 8.1 our title is relatively short. The title includes the crucial word "replication": "Proficiency mediates the effectiveness of corrective feedback: A close replication". This title contains **four main components** because, for us, these are the main components that define our study, and we want to convey these to our audience:

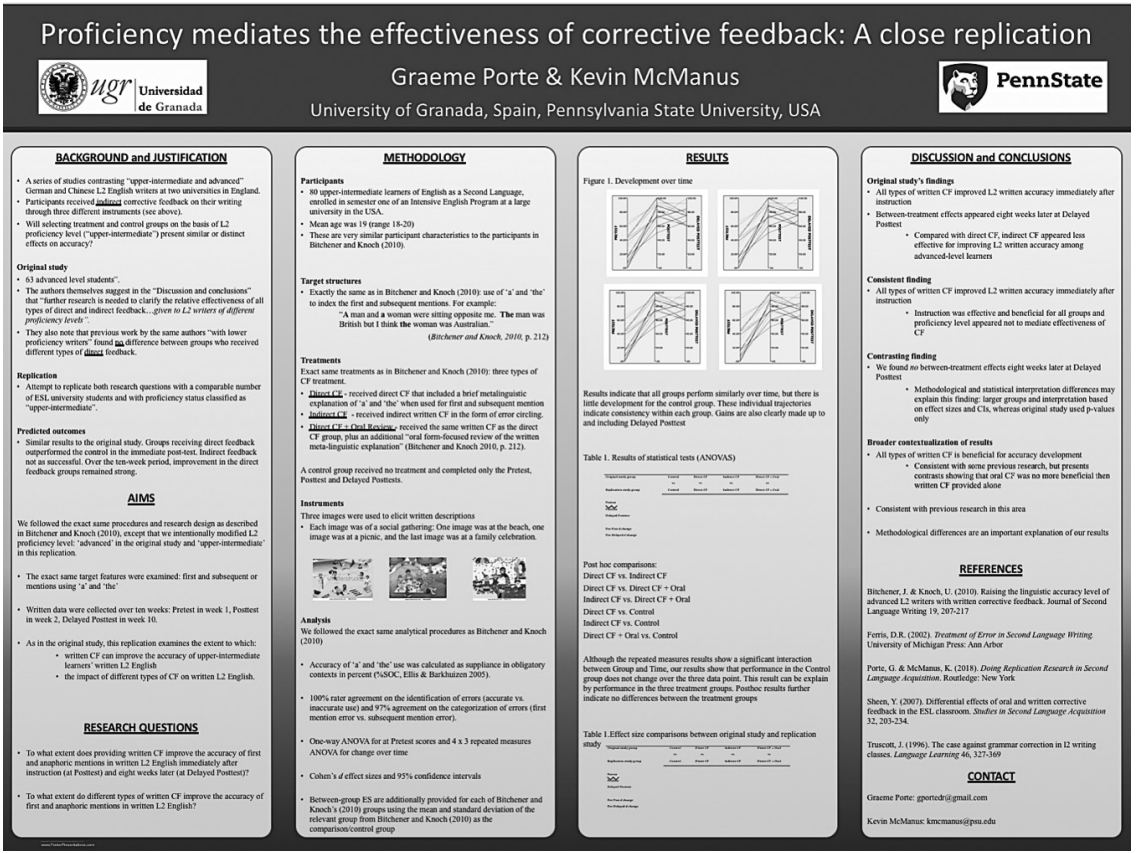


FIGURE 8.1 Example poster. To view this in colour and in closer detail please visit the eResource at www.routledge.com/9781138657359.

1. **Proficiency** was our intentional variable modification.
2. We examined the **impact of corrective feedback** on written accuracy.
3. Our variable modification and analyses allowed us to examine the **relative effectiveness of corrective feedback** at different proficiency levels by comparing our replication with the original.
4. Our study is a **replication**.

A helpful way to design your poster's title could be to boil down the study to three or four key words and then figure out how you can get these to work together in a title.

Next are the authors' names and affiliations. That should be relatively straightforward. Small images of our university logos on either side of authors' names provide a useful visual identity of affiliation. You will sometimes see that some posters include email addresses in this top part too, but, in our case, we have a bottom corner with that information. Decisions like those will come down to personal preference.

We are now going to work from left to right in filling our poster with content. Our first decision will be to rename the section titles and maybe resize them. How you use the space is up to you, and it is going to be influenced by the nature of your replication study and the messages you want to communicate.

Background and Aims

In "Background and justifications", we want to provide the essential information about (a) the original study, (b) our justification for replication, and (c) our variable modifications. It is also important to have essential and non-exhaustive information here, perhaps using bullet points if that will work. At any rate, we want to avoid blocks of text. The justification for replication may well be the bigger section here and will be based on your original rationale for replication.

In "Aims", we take a similar approach: we highlight in a short piece of text (a) that we very closely followed the same procedures and research design as in the original study and (b) our intentional variable modification was proficiency level, restating that the original study examined "advanced learners", but our replication examined "upper-intermediate" learners. We then boiled down our aims, as reported in Chapter 5, into three main bullet points:

- The same target features were examined, including what these were.
- The research design: ten weeks, pre-test, post-test, and delayed post-test.
- Restate research aims: (1) effects of written CF on L2 writing, and (2) differences between different types of CF.

You will notice that we had to be selective in what we reported because we want to avoid too much detail. We have to remember that the poster presentation

is a visualization. Our challenge is to decide on three or four essential aspects of the study's aims, and boil these down into short bullet points so that the audience can understand very quickly what we did. As discussed in Chapters 6 and 7, we also want to include phrases like "as in the original study", "exact same", and "very closely" to regularly remind our audience that we followed the original study as closely as possible.

Research Questions

We decided to include a separate section titled "Research questions" to make them stand out. This is a personal preference and they could be included under "Aims" if you are looking to save space. We used the same research questions as presented in Chapter 6.

Methodology

Our "Methodology" section is a full column, but it could be shorter depending on the nature of the replication study. For example, an intervention study, as in B&K (2010), requires description of the target structures and the treatments, which can make the methodology section a little longer when compared to studies that are not interventions.

» Activity 50: BOILING DOWN THE METHODOLOGY

The methodology is naturally a large section, and in a journal article it is going to cover multiple pages. Our challenge in creating a poster presentation is to provide the essential details that are necessary to understand our replication.

For each subsection, what are the essential details that are required to understand the replication?

- Participants:
    - What do we know about our participants' ages, backgrounds, prior learning experiences, proficiency level, courses they were enrolled in?
- Target structures:
    - What aspects of the target structure descriptions will you include and what will you leave out? What is your rationale for inclusion and exclusion?

- Treatments:
    - What aspects of the treatments are the essential points, do you think? How would you boil down the main information into a series of bullet points?
- Instruments:
    - It could be easier to provide examples of your instruments, rather than provide extensive details. What essential information about the instruments would you include?
- Analysis:
    - We spent a lot of time in Chapter 7 critiquing B&K's analysis and the writing up of our own. We can't include everything, and so the essentials are required. What would you consider essential for reporting our replication's analysis?

For our poster's methodology, we have five subsections: "Participants", "Target structures", "Treatments", "Instruments", and "Analysis". We will review what is in each of these sections. Your notes from Chapters 6 and 7 will doubtless contain lots of details about the methodology, but we will again have to be selective about what we report. Our guiding principle will be: include the basic information needed for an audience member to understand our replication. That principle will guide what we include and how we report our methodology.

For "Participants", we provided three bullet points. First, the number of participants, their proficiency level (i.e., upper-intermediate), that they were all ESL learners. Second, mean age, including range. Third, an explicit statement that these participant characteristics are "very similar" to those in the original study.

Our description of "Target structures" is brief: one bullet point and an example. We state that the target features are the same as in the original study: use of "a" and "the' for first and subsequent mentions, followed by an example (from the original study) illustrating use of these structures.

In "Treatments", we state that we used the same treatments as in the original study, with a very brief description of each treatment, followed by a description of the control group:

- *Direct CF* – metalinguistic explanation;
- *Indirect CF* – error circling;
- *Direct CF + Oral Review* – metalinguistic explanation plus oral review.

A control group received no treatment and completed only the pre-test, post-test and delayed post-tests.

You will notice that our "Instruments" includes reproductions of the three images used to elicit written descriptions, with a very brief statement that these were used to elicit written descriptions.

Last in this section is our "Analysis", which as you will remember had to be very detailed for our write up in Chapter 7. In the poster, however, as we need to cover the essential points, we will use a series of bullet points. As with our other sections, we state that we followed the same analytical procedures as in the original study. We then provide the following information:

- Construct operationalization: accuracy of "a" and "the" use was calculated as suppliance in obligatory contexts.
- Statistical tests used.
- Cohen's *d* effect sizes and 95% confidence intervals.

Results

Our results section contains perhaps the least amount of text compared with the other sections, with preference given to the more visually effective use of figures and tables. We draw on the following aspects from our Chapter 7 discussion of results: (1) data-accountable graphs showing individual scores over time per individual in each group, (2) descriptive results presented in a table, and (3) effect size comparisons between the original study and the replication to show similarities and differences.

Discussion and Conclusions

Last, in the far-right column, we begin with our "Discussion and conclusions", summarizing the main points discussed in Chapter 7. This section contains three subsections: "Consistent finding", "Contrasting finding", and "Broader contextualization of results". In these sections we summarize the crucial replication study's findings in light of the original study. In "Contrasting findings", in which we found no differences between the treatments at delayed post-test, we suggest that analytical differences could explain this finding: the original study's interpretation only used *p* values, whereas our use of effect sizes and confidence intervals provided clearer insights into the nature and magnitudes of between-group differences over time. Last in this section ("Broader contextualization of results"), we return to both the original study's motivation and our motivation for replication: the effects of CF on L2 written accuracy improvement. As discussed in Chapter 7, this is where we reengage with the original study's motivation.

Further Features and Considerations for the Poster Presentation

You will notice that we also include two small sections, one labelled "References" and the other labelled "Contact". You could also carve this space up differently to include, for example, an acknowledgements section.

Up to this point, we have walked through the planning and creation stages of a poster. Before we finish, we will offer just a few additional suggestions to help make your poster appealing and eye-grabbing, as well as what to expect at a poster session.

We didn't change any of the color features of our poster. One consideration to keep in mind is that bright colors, including colored text, may not always turn out as you planned. Sometimes it might be possible to get a practice print or a pre-print for you to check that the poster looks as you planned it. This will allow you to double-check any uses of color that you were unsure of. For example, maybe you had yellow backgrounds or red text in places, and so you want to check these display as planned.

You should also think about how you are going to transport your poster. A poster can be tricky to fly with and some printers offer the possibility of printing on fabric for easier transportation than flying with a poster tube. There is, of course, sometimes the possibility that you can print your poster on arrival at your destination.

When it comes to delivering your poster, you will find that you are given a poster board onto which you affix your poster. Hopefully you will be provided with pins, Velcro tabs, or tape to affix your poster, but you should check in advance and/or bring something with you just in case. You will then leave your poster displayed on the board until the poster session, allowing people to take a look in your absence. In the poster session itself, people will walk around all of the posters, stopping to take a look at each one, and maybe stopping longer at the ones that catch their interest. When somebody stops, don't be afraid to greet them, and then give them a little time to take in your poster. They may ask you questions, or you may offer a quick overview of your replication. When you start talking, you may find that your audience grows! All in all, enjoy the opportunity to talk to interested people about your study, your findings, and what this replication will lead to next.

Notes

1 www.nsf.gov (Publication NSF 18-053).
2 Kahneman, D. (2014). A new etiquette for replication. *Social Psychology*, 45(4), 310-311.
3 Neuliep, J., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, 5(4), 85-90.

9

EPILOGUE

Replication is both a much-neglected and much-needed feature of AL research. This need is not so much to enable us to generalize from previous studies but because AL research has typically sought to accumulate knowledge in a somewhat haphazard fashion – rather than organize and construct it by consciously building upon this previous work. The conclusion one might reasonably draw from such a *modus operandi* was that previous studies were not flawed to a sufficient degree, nor did they contain sufficient error that need worry us. We saw ourselves somehow sanctioned, therefore, to move on from earlier work and extend our related research into new contexts without further detailed attention to what had gone on previously.

But logic tells us that little, if any, experimental research contains no error whatsoever and that, before we even attempt to generalize *or* move on from it, we would be wise to appraise it closely.

Our argument is, if AL research is to be taken seriously as part of a truly contributory social science, its findings need to be minimally reliable and stand up to close scrutiny. If we do not make them so, we leave our research open to a less welcome kind of skepticism: are those research findings even reliable enough to find their way eventually into recommended teaching methodology, for example, or even into our language teaching textbooks?

The additional evidence we gather from our replication research will both accumulate knowledge and, crucially, help construct the knowledge we already have in a more useful way. Such construction of knowledge then becomes eminently contributory as supplementary evidence can potentially address a number of needs, including:

- Providing us with support for one hypothesis over another.
- Encouraging us to rework an original hypothesis.
- Suggesting the need to dismiss previous hypotheses.
- Helping reveal a flaw in what we previously assumed to be correct.
- Inspiring a *new* hypothesis or new research question.[1]

Consequently, our additional evidence will almost inevitably be inconclusive.

That last statement should recall the diagram of the research process you saw back in the Introduction (p. 2) and its essentially cyclical nature. Everything we research is inevitably and enticingly temporary: science is always a work in progress and – as Sagan reminded us in our *Introduction* – has therefore to be open to challenge and rectification as part of the striving for progress and the advancement of knowledge.

Note

1 Adapted from https://undsci.berkeley.edu/article/_0_0/howscienceworks_10.

INDEX